

Copyright  
by  
Eleisha Lynnette Jackson  
2016

The Dissertation Committee for Eleisha Lynnette Jackson  
certifies that this is the approved version of the following dissertation:

**Protein design, modeling, and the evolution of proteins**

Committee:

---

Claus O. Wilke, Supervisor

---

Nancy Moran

---

Hans Hofmann

---

Christopher Sullivan

---

Andrew Ellington

**Protein design, modeling, and the evolution of proteins**

**by**

**Eleisha Lynnette Jackson, B.S.**

**DISSERTATION**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**DOCTOR OF PHILOSOPHY**

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2016

To my friends and family

## Acknowledgments

I would like to thank my advisor Dr. Claus Wilke for his guidance throughout my graduate career. In addition, I also thank the members of my committee and the members of the Wilke Lab for their feedback and support during my tenure here at the University of Texas at Austin. Lastly, I would like to thank my family and friends for their emotional support throughout my graduate career.

# **Protein design, modeling, and the evolution of proteins**

Eleisha Lynnette Jackson, Ph.D.  
The University of Texas at Austin, 2016

Supervisor: Claus O. Wilke

Proteins are crucial players in the functional processes that allow for cellular life. Changes in the sequences of proteins have consequences for how these proteins function. Therefore, the study of how proteins change over time has been a central question in the field of evolutionary biology. As our understanding of how proteins function and change increases, we are not only able to test our hypotheses but we are also able to design and model new proteins, which is the ultimate test of our knowledge of how proteins function. Using the information from our protein modeling attempts, we can learn more about how natural proteins function and change over time. In this dissertation, I used protein modeling techniques to understand protein evolution. In Chapter 2, I assessed how closely designed proteins recapitulate observed patterns in natural proteins. I have found that designing proteins with a flexible-backbone protocol results in site variability that more closely mimics what is seen in natural proteins. In addition, I have also found that, in designed proteins, hydrophobic residues are often underrepresented in the core of the protein.

These results suggest that our scoring functions and/or backbone sampling methods could be further improved. In Chapter 3, I used protein design to predict site-wise evolutionary rates in proteins. I found that protein design is a poor predictor of evolutionary rate, explaining only approximately  $\sim 7\%$  of the variation in rate across sites in enzymes. In Chapter 4, I used protein design and homology modeling to predict tolerance to deletions in enhanced green fluorescent protein. I also compared these predictions to predictions made using other structural properties including solvent accessibility, local packing density and secondary structure. I found that when combining computational scores from modeled structures along with other structural properties (i.e., local packing density, solvent accessibility and secondary structure) as predictors, I was largely able predict whether or not a given deletion would result in a functional protein product. Finally, in Chapter 5, I developed a computational pipeline to assess binding affinity in protein-protein interactions. I used this pipeline to recapitulate patterns of Machupo virus entry across various species. Taken together, the work presented in this dissertation has given us insight into which structural constraints affect protein evolution.

# Table of Contents

<b>Acknowledgments</b>	<b>v</b>
<b>Abstract</b>	<b>vi</b>
<b>List of Tables</b>	<b>xi</b>
<b>List of Figures</b>	<b>xii</b>
<b>Chapter 1. Introduction</b>	<b>1</b>
1.1 Figures . . . . .	6
<b>Chapter 2. Amino-acid site variability among natural and designed proteins</b>	<b>7</b>
2.1 Introduction . . . . .	7
2.2 Materials and Methods . . . . .	9
2.2.1 Data sets . . . . .	9
2.2.2 Protein design . . . . .	10
2.2.3 Data analysis . . . . .	11
2.3 Results . . . . .	13
2.3.1 Overall site variability . . . . .	14
2.3.2 Amino-acid distributions . . . . .	16
2.3.3 Site variability and solvent accessibility . . . . .	18
2.4 Discussion . . . . .	21
2.5 Figures . . . . .	28
<b>Chapter 3. Intermediate divergence levels maximize the strength of structure–sequence correlations in enzymes and viral proteins</b>	<b>43</b>
3.1 Introduction . . . . .	43
3.2 Materials and methods . . . . .	46



3.2.1	Structures, sequences, and measures of sequence properties	46
3.2.2	Protein Design . . . . .	49
3.2.3	Calculation of structural properties . . . . .	51
3.3	Results . . . . .	52
3.3.1	Structural Predictors of Evolutionary Rate . . . . .	53
3.3.2	Protein Design as a Structural Predictor . . . . .	54
3.3.3	Effect of Divergence of Structure–Rate Relationships . .	56
3.4	Discussion . . . . .	61
3.5	Figures . . . . .	67
3.6	Tables . . . . .	80
<b>Chapter 4.</b>	<b>Computational prediction of the tolerance to amino-acid deletion in green-fluorescent protein</b>	<b>81</b>
4.1	Introduction . . . . .	81
4.2	Materials and Methods . . . . .	84
4.2.1	Functional Data for Mutants . . . . .	84
4.2.2	Calculation of Structural Properties . . . . .	85
4.2.3	Structural Modeling . . . . .	86
4.2.4	Statistical Analysis of Functional Status . . . . .	89
4.3	Results . . . . .	91
4.3.1	Variation in structural properties between non–tolerated and tolerated deletions . . . . .	92
4.3.2	Functional Classification Prediction . . . . .	92
4.4	Discussion . . . . .	96
4.5	Figures . . . . .	99
4.6	Tables . . . . .	103
<b>Chapter 5.</b>	<b>Computational prediction of zoonotic transmission of Machupo Virus</b>	<b>106</b>
5.1	Introduction . . . . .	106
5.2	Materials and Methods . . . . .	108
5.3	Results . . . . .	112
5.4	Discussion . . . . .	114
5.5	Figures . . . . .	117

<b>Chapter 6. Conclusion</b>	<b>122</b>
6.1 Discussion . . . . .	122
<b>Bibliography</b>	<b>126</b>
<b>Vita</b>	<b>147</b>

## List of Tables

3.1	Averages of Spearman correlation coefficients between structural properties and evolutionary rate (ER). The structural properties analyzed are RSA, WCN, and predicted rate of designed proteins (DR). The analysis was performed on two data sets, one comprised of 208 enzyme monomers and comprised of nine viral proteins. Structure-ER correlations are higher in absolute magnitude in enzymes. . . . .	80
4.1	Summary of AUC values for logistic regression models using structure to predict functional status. The structural properties analyzed are RSA, WCN, SS, and mean score. The SS of a residue was classified as beta sheet, alpha helix or loop. Mean score is the mean of the Rosetta scores for 100 models of a given mutant. Each property was used as a single predictor or in combination with the other three structural predictors to predict the functional status of a given mutant. Functional status was taken from Arpino <i>et al.</i> [5]. We report the mean Area Under the Curve (AUC) of 100 trials for each model for the test data (mean cross-validated AUC). We also report the AUC for the model fitted on the entire data set (AUC of Model). Models are sorted in decreasing order by mean cross-validated AUC. The model with RSA, WCN and mean score has the best predictive ability. . . . .	104
4.2	Summary of AUC values when using a support vector machine to predict functional status. The structural properties analyzed are RSA, WCN, SS, and mean score. Mean score is the mean of the Rosetta scores for 100 models of a given mutant. Each property was used as a single predictor or in combination with the other three structural predictors to predict the functional status of a given mutant. We report the mean Area Under the Curve (AUC) of 100 trials for each model for the test data (cross-validated AUC). We also report the AUC of the model for the model fitted on the entire data set (Model AUC). Models are sorted in decreasing order by mean cross-validated AUC. The model that is the best at making predictions is the model with RSA, WCN and mean score as structural predictors. . . .	105

# List of Figures

1.1	Description of Structural Properties. (A) Visualization of Solvent Accessibility. (B) Visualization of Local Packing Density. Each colored red particle represents a residue in the protein. In A, the lower red particle represents a surface residue. The red and white molecules indicate solvent molecules (e.g., water) that are contacting the red amino acid. This residue has a larger solvent accessibility because there is a larger proportion of the residue surface exposed to solvent. The upper red particle represents a core residue. This residue is not in contact with any solvent molecules and thus has low solvent accessibility. Relative solvent accessibility is obtained by normalizing the solvent accessibility of a given residue by the maximum amount of solvent accessibility for that amino acid. In B, the arrows pointing towards each residue indicate contacts between the red focal residue and its neighboring residues. The upper red residue represents a residue that has many neighbors (represented by the arrows) and thus has a high weighted contact number. The lower red residue is a surface amino acid with few neighbors and thus has a lower weighted contact number. . . .	6
2.1	Mean site entropy for designed and natural proteins. Each box-plot represents the distribution of mean site entropies within the respective dataset (A: yeast proteins; B: protein domains). “FB” refers to fixed-backbone design. Temperature values refer to the design temperature used during the Backrub design method. “NS” refers to natural sequences. “Soft” refers to the Soft design method. We find generally that increased backbone flexibility allows for more site variability. Intermediate temperatures produce site variabilities most similar to those seen in natural sequences. Overall, natural sequences in the protein-domains data set are more variable than are those in the yeast-proteins data set. . . . .	28
2.2	Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the yeast-proteins data set. For designed proteins, only flexible-backbone designs with design temperature 0.6 were considered. (A) Overall frequencies. (B) frequencies at exposed sites (defined as sites with $RSA > 0.05$ ). (C) frequencies at buried sites (defined as sites with $RSA \leq 0.05$ ). . . . .	29

2.3	Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the yeast-proteins data set. For designed proteins, only fixed-backbone designs were considered. (A) overall frequencies. (B) frequencies at exposed sites (defined as sites with $\text{RSA} > 0.05$ ). (C) frequencies at buried sites (defined as sites with $\text{RSA} \leq 0.05$ )	30
2.4	Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the yeast-proteins data set. For designed proteins, only flexible-backbone designs with design temperature 1.2 were considered. (A) overall frequencies. (B) frequencies at exposed sites (defined as sites with $\text{RSA} > 0.05$ ). (C) frequencies at buried sites (defined as sites with $\text{RSA} \leq 0.05$ ). . . . .	31
2.5	Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the protein-domains data set. For designed proteins, only fixed-backbone designs were considered. (A) overall frequencies. (B) frequencies at exposed sites (defined as sites with $\text{RSA} > 0.05$ ). (C) frequencies at buried sites (defined as sites with $\text{RSA} \leq 0.05$ )	32
2.6	Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the protein-domains data set. For designed proteins, only flexible-backbone designs with design temperature 0.6 were considered. (A) overall frequencies. (B) frequencies at exposed sites (defined as sites with $\text{RSA} > 0.05$ ). (C) frequencies at buried sites (defined as sites with $\text{RSA} \leq 0.05$ ) . . . . .	33
2.7	Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the protein-domains data set. For designed proteins, only flexible-backbone designs with design temperature 1.2 were considered. (A) overall frequencies. (B) frequencies at exposed sites (defined as sites with $\text{RSA} > 0.05$ ). (C) frequencies at buried sites (defined as sites with $\text{RSA} \leq 0.05$ ) . . . . .	34

2.8	Mean Kullback-Leibler (KL) divergence for designed and natural proteins, shown for the protein-domain data set. A higher KL divergence indicates that the amino-acid distributions at sites in designed proteins are less similar to the corresponding distributions in the natural proteins. “FB” refers to fixed backbone design and “NS” refers to the control case where natural sequences are compared to themselves. (A) KL divergence calculated from the relative frequencies of the 20 amino acids. (B) KL divergence calculated from rank-ordered frequency distributions. The most common amino acid in the reference distribution is compared to the most common amino acid in the focal distribution, the same is done for the second-most common amino acid, and so on, irrespective of the type of amino acids. . . . .	35
2.9	Mean Kullback-Leibler (KL) divergence for designed and natural proteins, shown for the yeast-proteins data set. A higher KL divergence indicates that the amino-acid distributions at sites in designed proteins are less similar to the corresponding distributions in the natural proteins. “FB” refers to fixed backbone design, and “NS” refers to the control case where natural sequences are compared to themselves. (A) KL divergence calculated from the relative frequencies of the 20 amino acids. (B) KL divergence calculated from rank-ordered frequency distributions. The most common amino acid in the reference distribution is compared to the most common amino acid in the focal distribution, the same is done for the second-most common amino acid, and so on, irrespective of the type of amino acids. . . . .	36
2.10	Site entropy versus Relative Solvent Accessibility (RSA) for designed and natural sequence alignments of the protein S-formylglutathione hydrolase (PDB: 1PV1, chain A). Natural sequences exhibit a clear trend of higher site variability at higher RSA values. The flexible backbone designs exhibit a similar trend but the fixed backbone designs do not. . . . .	37
2.11	Distributions of correlation coefficients between site entropy and RSA, for the protein-domain data set. “FB” indicates fixed-backbone design and “NS” indicates natural sequences. (A) Distributions represented as boxplots. (B) Correlation coefficients for individual proteins. Lines connect identical structures in the different design conditions. The color shading represents the strength of the correlation for the natural sequence alignment. In general, natural proteins display a stronger correlation between site entropy and RSA than designed proteins. . . . .	38

2.12	Distributions of correlation coefficients between site entropy and RSA, for the yeast-proteins data set. “FB” indicates fixed-backbone design, “Soft” indicates soft backbone design, and “NS” indicates natural sequences. (A) Distributions represented as boxplots. (B) Correlation coefficients for individual proteins. Lines connect identical structures in the different design conditions. The color shading represents the strength of the correlation for the natural sequence alignment. In general, natural proteins display a stronger correlation between site entropy and RSA than designed proteins. . . . .	39
2.13	Median of the distribution of mean sequence entropies for designed and natural sequences, calculated separately for buried (black), partially buried (blue), and exposed (red) residues (A: yeast proteins; B: protein domains). We defined buried sites as those with $RSA \leq 0.05$ , partially buried as those with $0.05 < RSA \leq 0.25$ , and exposed as those with $RSA > 0.25$ . Dashed lines indicate the corresponding median for natural sequence alignments. Note that for buried (black) and partially buried (blue) residues, the temperatures at which natural site variability and design variability match are comparable. By contrast, for exposed residues, a higher design temperature is required for the design variability to match the natural site variability.	40
2.14	Distribution of correlation coefficients between RSA and site entropy for hybrid designs and for natural proteins (A: yeast proteins; B: protein domains). “FB” indicates fixed-backbone design and “NS” indicates natural sequences. For the hybrid designs, buried and partially buried sites were taken from sequences designed at one temperature, and exposed sites were taken from sequences designed at a different temperature. For the hybrid designs, the correlation coefficients were similar to those of natural sequences (paired $t$ test, $P = 0.517$ ( $T = \text{FB}, 0.1$ ) and $P = 6.78 \times 10^{-8}$ ( $T = 0.03, 0.1$ ) [yeast proteins], $P = 5.19 \times 10^{-5}$ ( $T = 0.3, 1.8$ ) and $P = 0.118$ ( $T = 0.6, 1.8$ ) [protein domains]). . . . .	41
2.15	Correlation coefficients between RSA and site entropy for hybrid designs and natural proteins. For the hybrid designs, buried and partially buried sites were taken from proteins designed with a fixed backbone (yeast proteins) or a temperature of $T = 0.6$ (protein domains). Exposed residues were taken from proteins designed with a temperature of $T = 0.1$ (yeast proteins) or $T = 1.8$ (protein domains). The solid line indicates $y = x$ . Note that while the range of correlation values in hybrid designs generally matches the range of values in natural alignments, predictions for specific proteins are not that accurate. . . . .	42

3.1	Distribution of correlation coefficients between structural properties and evolutionary rate (ER). (A) Spearman correlation coefficients between RSA and ER for the two data sets ( $t$ test: $P = 3.324 \times 10^{-5}$ ). (B) Spearman correlation coefficients between WCN and ER for the two data sets. For all structural properties, on average, viral proteins show weaker correlations than do enzymes ( $t$ test: $P = 2.454 \times 10^{-5}$ ). . . . .	67
3.2	Comparison of structure–rate correlations for the full data set of enzymes and the designed set. (A) Comparison of Spearman correlation coefficients for WCN–ER. (B) Comparison of Spearman correlation coefficients for RSA–ER. For both WCN–ER and RSA–ER the mean of the distributions for the designed set of enzymes is the same as that of the full data set of enzymes ( $t$ test: $P = 0.947$ for WCN–ER, $P = 0.419$ for RSA–ER). . .	68
3.3	Correlation Coefficients of Design Rate and evolutionary rate (ER). Distributions of Spearman correlation coefficients between design rate (DR) and evolutionary rate (ER) for the two data sets. Enzyme proteins have higher correlations on average ( $t$ test: $P = 7.50 \times 10^{-4}$ ). . . . .	69
3.4	Distribution of $R^2$ for linear models of structural predictors of evolutionary rate (ER) in enzymes. WCN, RSA, DR and all combinations were used as predictors in a linear model with ER at sites as the response. Very little variation in ER can be explained when using design rate (DR) as a single predictor. For enzymes, only 32 proteins were included . . . . .	70
3.5	Distribution of $R^2$ for linear models of structural predictors of ER in viruses. WCN, RSA, DR and all combinations were used as predictors in a linear model with evolutionary rate at sites as the response. Very little variation in evolutionary rate can be predicted by RSA, WCN or DR in viral proteins. . . . .	71
3.6	Distribution of average structural properties for each protein in the two data sets. (A) Distribution of average RSA. The distribution of average RSA different are very similar for both data sets ( $t$ test: $P = 0.027$ ). (B) Distribution of average WCN. The distribution of average WCN is the same for both data sets ( $t$ test: $P = 0.437$ ). . . . .	72



3.7	Comparison of structure–rate correlations with Mean RSA. (A) Spearman correlations of WCN–ER vs. mean RSA. Proteins with residues that are more exposed on average have slightly larger WCN–ER correlations in magnitude (Spearman’s correlation test: $\rho = 0.181$ , $P = 7.653 \times 10^{-3}$ ). (B) Correlations of RSA–ER vs. mean RSA. Proteins with residues that are more exposed on average also have slightly larger RSA–ER correlations in magnitude (Spearman correlation test: $\rho = -0.228$ , $P = 7.241 \times 10^{-3}$ ). . . . .	73
3.8	Comparison of structure–rate correlations with Mean WCN. (A) Spearman correlations of WCN–ER vs. mean WCN (Spearman correlation test: $\rho = -0.082$ , $P = 0.2283$ ). (B) Correlations of RSA–ER vs. mean WCN (Spearman correlation test: $\rho = 0.077$ , $P = 0.2585$ ). The average WCN of a protein is not related to the strength of structure–rate correlations. . . . .	74
3.9	Divergence of sequences within the data sets. (A) Distributions of mean patristic distances for sequences in each protein alignment. Enzymes have larger mean patristic distances ( $t$ test: $P < 2.2 \times 10^{-16}$ ). (B) Distributions of mean root-to-tip distances for sequences in each protein alignment. Enzymes have larger mean root-to-tip distances ( $t$ test: $P < 2.2 \times 10^{-16}$ ). For both measures of divergence, the proteins within the enzyme dataset are more diverged. Divergence is relatively low between the viral proteins. . . . .	75
3.10	Comparison of the mean of entropy and the variance of entropy for individual proteins. (A) Variance in entropy at sites compared against overall mean entropy for each protein. Five different enzymes are highlighted, spanning the range of different combinations of high and low mean entropy and entropy variance. The enzymes are colored in black and the virus proteins are colored red. (B)–(F) Distributions of site-wise entropy values for the five proteins highlighted in A. There are a variety of distributions in site entropy for different proteins. Note: The protein denoted by the PDB ID 3GOL is a viral protein. . . .	76

3.11	Comparison of structure–rate correlations with variance of entropy at sites. (A) Comparison of Spearman Correlation Coefficients of WCN–ER and variance of entropy for proteins. (Spearman’s correlation test: $\rho = -0.321$ , $P = 1.526 \times 10^{-6}$ using only the original protein data sets) (B) Correlations of RSA–ER and variance of entropy for proteins ( $\rho = 0.236$ , $P = 4.756 \times 10^{-4}$ using only the original protein data sets). Enzymes are black, the viral proteins with the original alignments are in red, and the viral proteins with the newly collected sequences are in turquoise. Enzymes have more variance in entropy across proteins and have larger structure–rate correlations in magnitude for both RSA and WCN. Virus proteins represented by the newly curated, more diverged alignments (see Methods) have similar structure–rate correlations to the original viral protein data set. . . . .	77
3.12	Comparison of structure–rate correlations with mean Guidance scores of proteins. (A) Comparison of Spearman correlation coefficients for WCN–ER. (B) Comparison of Spearman correlation coefficients for RSA–ER. Enzymes are black and viral proteins in red. Enzymes have more variation in alignment quality among proteins and have a non-significant relationship between alignment quality and structure–rate correlations (Spearman’s Correlation test: $\rho = -0.023$ , $P = 0.746$ for WCN–ER and $\rho = -0.132$ , $P = 0.057$ for RSA–ER). For viral proteins there is no significant relationship between alignment quality and structure–rate correlations ( $\rho = -0.633$ , $P = 0.076$ for WCN–ER and $\rho = 0.317$ , $P = 0.410$ for RSA–ER). . . . .	78
3.13	Comparison of structure–rate correlations with divergence. (A) Spearman correlations of WCN and ER vs. mean pairwise distance (Spearman’s correlation test: $\rho = -0.117$ , $P = 0.086$ for WCN–ER). (B) Correlations of RSA and ER vs. mean pairwise distance. Enzymes are black, the viral proteins with the original alignments are in red, and the viral proteins with the newly collected sequences are in turquoise. Proteins that are more diverged (as represented by mean pairwise distance) have stronger RSA–ER correlations (Spearman’s correlation test: $\rho = 0.161$ , $P = 0.017$ ). . . . .	79
4.1	Structural Representation of enhanced GFP (EGFP). Secondary structure elements are colored. Beta sheets are colored in red, alpha helices are colored in cyan and loops are in magenta. The chromophore responsible for florescence is colored in green. The structure of EGFP is formed by a beta-barrel composed of eleven beta sheets. The chromophore that results in the green fluorescent phenotype is in the middle of an alpha helix that is housed in the middle of this beta-barrel. . . . .	99

4.2	Visualization of the computational modeling pipeline. The colors represent variation between structural models produced by each protocol. After removing the chromophore, we used the relax protocol in Rosetta to optimize the structure. We chose the lowest scoring model from the 100 created models as our modeling template. We used RosettaRemodel to model the EGFP structure without the chromophore. Using the lowest scoring model from the protocol as our template, we used Modeller to model 25 mutants for each deletion mutant. We used the relax protocol to create four optimized structures for each of the 25 homology models for each mutant. We took the mean of the 100 models and used this as a predictor of functional status for each mutant. . . . .	100
4.3	Distributions of Structural Properties for EGFP Mutants. (A) Distribution of RSA. Residues with tolerated deletions are more exposed than residues with non-tolerated deletions ( $t$ test: $P = 1.030 \times 10^{-3}$ ). (B) Distribution of WCN. On average, residues with tolerated deletions have lower WCN than residues with non-tolerated deletions. ( $t$ test: $P = 2.998 \times 10^{-7}$ ). (C) Distribution of mean score for EGFP Mutants. Residues that are tolerant to deletion have lower scores (i.e., more negative) on average than non-tolerant residues ( $t$ test: $P = 2.084 \times 10^{-6}$ ). (D) Secondary structure of mutants. Non-tolerated deletions are colored in blue and tolerated deletions are in red. The majority of the residues deleted in the loop regions and alpha helix regions are tolerated and result in a functioning fluorescent phenotype. 78.3% and 66.7% of deleted residues are tolerated in loop and helical regions, respectively. However, only a small fraction of residues (21.6%) deleted in areas of the proteins that make up a beta sheet are tolerated. . . . .	101
4.4	Comparison of Data Along Principal Component 1 versus Principal Component 2. A) Plot of PC1 vs. PC2. Data points are colored according to functional status. Functional mutants are blue and non-functional mutants are in red. Mutants are largely separated along PC1. B) Loadings of structural properties along principal component axes PC1 and PC2. Most structural properties contribute to PC1 except for beta sheet with is mostly loaded onto PC2. . . . .	102
4.5	Comparisons of mean cross-validated AUC from SVM and Logistic Regression Models. For each model that has the exact same predictors the cross-validated AUC value from the SVM is plotted against the cross-validated AUC value from the logistic model. The dotted gray line represents the line $y = x$ . For all but one model, logistic regression models with the same predictors have higher mean cross-validated AUC values. . . .	103

5.1	Alignment of TfR1 receptors from various species along with the tested chimeras. The sequence numbering corresponds to the amino-acid position in the hTfR1 sequence. The five naturally occurring receptors included in this study are: <i>R. norvegicus</i> , <i>M. musculus</i> , <i>C. callosus</i> (the native host of Machupo virus), <i>H. sapiens</i> , and the <i>H. sapiens</i> L212V variant. The rat-short chimera is the rat TfR1 with a five residue swap from <i>C. callosus</i> which is indicated in red. The rat-long chimera is the rat TfR1 with a ten residue swap which includes the five residues from rat-long along with five additional amino acids (colored in blue). The mouse-human chimera is the mouse TfR1 with a five amino acid swap from the human TfR1. These amino acids are colored in green. The human L212V is identical to the hTfR1 except that there is a valine at position 212. This valine is colored in magenta in the hTfR1 L212V sequence. . . . .	117
5.2	Interaction between human transferrin receptor 1 (hTfR1) and the Machupo Virus Glycoprotein 1 (MACV GP1). hTR1 is colored blue and MACV GP1 is colored blue. This is a visualization of the interaction between the apical domain of human TfR1 and MACV GP1. MACV uses its GP1 to bind to the TfR1 on the surface of the host's cell by interacting with the apical domain of TfR1. . . . .	118
5.3	Computational Pipeline Overview. For each MACV GP1-TfR1, the target TfR1 sequence was aligned to the hTfR1 structure before modeling. After modeling each protein complex in Modeller, the complexes were re-docked using RosettaDock. Convergence of the docking protocol was assessed by plotting a RMS versus Interface Score plot and checking for a funnel. The mean interface score for the top ten scoring models for each MACV GP1-TfR1 complex was used as the proxy for binding affinity in subsequent analyses. . . . .	119
5.4	Mean Interface Scores for modeled MACV GP1-TfR1 Complexes. Each boxplot represents the distribution of the top ten scoring models for each TfR1 by interface score. Each boxplot is colored according to known infectivity information. Green coloring indicates efficient TfR1 receptors for entry. Yellow indicates receptors that are partially efficient. Red coloring indicates receptors that cannot be used as efficient receptors for entry. Overall, inefficient receptors have less negative interface scores indicating that binding is not as effective in those models. The human L212V model also has a much less negative average interface score as compared to the human model. This is consistent with experimental results suggesting that this SNP provides some protection from MACV <i>in vitro</i> . . . . .	120

5.5	Interface between MACV GP1 and hTfR1. (A) Complex between MACV GP1-hTfR1 L212V. MACV GP1 is colored in magenta and the hTfR1 is colored blue. The L212V mutant is colored in green. (B) Superposition of a modeled JUNV GP1 onto the MACV GP1. JUNV GP1 is colored in cyan and MACV GP1 is colored in magenta. The loop region unique to MACV is colored pink. The JUNV GP1 is rotated relative to the position of the MACV GP1. (C) Alignment of the GP1 sequences of Machupo, Junín, Sabia, Chapare, and Guanarito. Red asterisks highlight residues that contact hTfR1 in MACV. Interaction residues between hTfR1 and MACV GP1 were calculated using FoldX [86]. . . . .	121
-----	--	-----

# Chapter 1

## Introduction

Proteins are the workhorses of the body. Proteins have essential roles in important biological processes such as catalysis and immune system function. Changes, such as amino acid substitutions, within the sequence of a protein can affect how the protein functions. The evolutionary rate of a protein describes the rate at which its sequence changes over time. Thus, understanding the evolutionary rate of proteins is critical to our understanding of biological processes and how they develop over time. There have been numerous studies to elucidate the factors that impact the evolutionary rate of proteins. At the whole protein level, protein expression has been shown to be the strongest predictor of evolutionary rate [22]. Proteins that are highly expressed are more conserved and thus evolve more slowly. Other factors that have impacted protein evolution include interactions with other proteins [33,71,110] and selection against translational errors [23].

However, within a given protein, sites may have widely different rates of evolution. For example, active sites in enzymes are more conserved due to functional importance. In addition, sites within viral proteins that are responsible for viral entry might show signatures of rapid evolution due the

viral-host arms race that many of these sites participate in [92]. In summary, within a given protein there may be sites that are highly conserved and sites that show signatures of rapid evolution.

Many of the factors that have been found to constrain the rates at which protein sites evolve are biophysical in nature [56,90,109]. Among these biophysical constraints, structural constraints in particular have emerged as key constraints of evolutionary rate at sites. Two of the most important structural constraints on protein evolution constraints are solvent accessibility and local packing density. A residue's solvent accessibility is commonly measured by its relative solvent accessibility (RSA) (Figure 1.1A). RSA measures how much solvent (ex. water) a given residue is exposed to. Residues that have high solvent accessibility (i.e., have high exposure to solvent) are less conserved, exhibit more sequence variability and evolve more quickly [12, 31, 32, 38, 65, 69, 82, 85]. Local packing density (LPD) measures how densely packed a residue is among its neighbors. A residue's LPD is often measured by its weighed contact number (WCN) (Figure 1.1B). Residues that are densely packed are more conserved, and evolve more slowly as compared to sites with fewer contacts [42,55,88,111].

At the same time that this research on the evolution of proteins has expanded, there have also been several advances in our ability to model and design proteins. The area of protein design focuses on finding low energy sequences that are compatible with a given structure. There have been several methods that been developed to design proteins. These methods include both

deterministic methods such as dead-end elimination [21,36,37,40] and stochastic methods [51,53]. The applications of these methods have expanded as we have developed the capability to dock proteins and hence study protein-protein interactions [14,16,39,74].

Although the fields of protein design and protein evolution were originally separate, recently, the field of protein design has lead to several important discoveries in the field of protein evolution. Likewise, the study of natural proteins has helped discover some possible improvements that can be made to our algorithms to improve design. As both fields (protein design and protein evolution) learn more, there will be numerous new discoveries at the intersection of these two fields. This thesis is comprised of various studies at the intersection of protein design and protein evolution. These studies represent part of a developing bridge between these two fields by using the techniques developed in the area of protein design to better understand the evolution of natural proteins.

The second and third chapters of this thesis use methods of protein design to predict substitutions at sites in natural proteins. In the second chapter, I compared designed proteins to natural proteins to understand how designed proteins mimic natural proteins and how they differ. I found that designed proteins do not accurately recapitulate the relationship between site variability and solvent exposure in proteins. Additionally, I found that, on average, hydrophobic amino-acids are underrepresented in the core of designed proteins. In the third chapter, I used protein design as a structural predic-



tor of evolutionary rate in proteins. Because protein design seeks to find low energy sequences for a given structure, to some extent, the sequences found using protein design can be representative of structural constraints on a protein's sequence. I compared the ability of protein design to predict site-wise evolutionary rate with the ability of two other prominent predictors, WCN and RSA. I found that both RSA and WCN are much stronger predictors of evolutionary rate at sites. In fact, protein design on its own was found to be a poor predictor of evolutionary rate at sites explaining at most only approximately  $\sim 7\%$  of the variance in site-wise evolutionary rates across a protein. In addition, I found that divergence within alignments used to calculate evolutionary rates for a protein had an impact on the strength of structure-rate correlations.

In the fourth chapter, I used protein design and homology modeling techniques to predict the effect of deletions on protein function. Using enhanced green fluorescent protein (EGFP) as a model protein, I used protein design and homology modeling to explicitly model amino-acid deletions in EGFP. I then used machine learning techniques to use the computational scores from my modeling techniques to predict the functional status of a deletion. I found that protein modeling can be used to predict whether or not a given deletion will still result in a functional protein product. However, protein modeling was not more predictive of function when compared to WCN, a simpler structural property. Even so, when combining protein modeling in a predictive model and other structural predictors (ex. WCN, RSA, secondary structure), I observed

an increase in predictive power. Therefore I found that protein modeling can be important for building more accurate, predictive models of deletions.

In the fifth and final chapter, I used homology modeling and protein-protein docking to predict how mutations affect binding in protein-protein interactions. I developed a computational pipeline using both protein homology modeling and protein-protein docking tools to predict the effect of amino-acid substitutions on binding affinity. Using this pipeline, I managed to recapitulate the host-virus binding, and hence entry, patterns observed in Machupo virus, a New World Arenavirus.

These projects represent only a section of the work done in the broad area of research at the intersection of protein design and evolution. Improvements in our abilities to design and model proteins will allow us to use designed proteins to perform accurate computational studies of proteins. Likewise, a better understanding of the properties of natural proteins and how their sequences are shaped over time by structure will provide insightful information that will be critical for the development of more accurate energy functions and more complete methods for searching sequence space. These developments will help improve our ability to model proteins and protein-protein interactions.

## 1.1 Figures

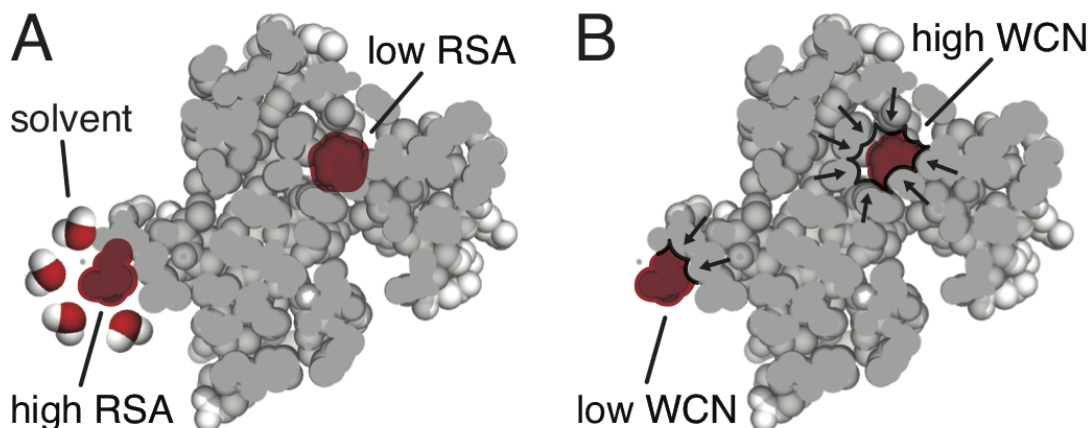


Figure 1.1: Description of Structural Properties. (A) Visualization of Solvent Accessibility. (B) Visualization of Local Packing Density. Each colored red particle represents a residue in the protein. In A, the lower red particle represents a surface residue. The red and white molecules indicate solvent molecules (e.g., water) that are contacting the red amino acid. This residue has a larger solvent accessibility because there is a larger proportion of the residue surface exposed to solvent. The upper red particle represents a core residue. This residue is not in contact with any solvent molecules and thus has low solvent accessibility. Relative solvent accessibility is obtained by normalizing the solvent accessibility of a given residue by the maximum amount of solvent accessibility for that amino acid. In B, the arrows pointing towards each residue indicate contacts between the red focal residue and its neighboring residues. The upper red residue represents a residue that has many neighbors (represented by the arrows) and thus has a high weighted contact number. The lower red residue is a surface amino acid with few neighbors and thus has a lower weighted contact number.

## Chapter 2

# Amino-acid site variability among natural and designed proteins

### 2.1 Introduction

This work has been previously published in the journal *PeerJ*.<sup>1</sup>

Computational protein design has made tremendous progress in recent years. For example, computational design has been used successfully to engineer proteins that bind to an influenza virus [29], to create enzymes [83], and to develop novel protein folds not seen in nature [51]. All these examples have in common that many different computational predictions were generated, and among the best were a few that worked experimentally. Thus, while computational design can produce specific sequences that fold correctly and are functional, it is much less clear how similar designed proteins are on average to natural proteins of a comparable fold.

There are several patterns of sequence variation that are consistently seen in natural proteins. For example, amino acid frequencies follow charac-

---

<sup>1</sup>E. L. Jackson, N. Ollikainen, A. W. Covert III, T. Kortemme and C. O. Wilke. Amino-acid site variability among natural and designed proteins. *PeerJ*, e:211, 2013. N. Ollikainen and A. W. Covert III helped perform the experiments. N. Ollikainen helped design and perform the experiments and write the paper. T. Kortemme and C. O. Wilke helped to design the project and write the manuscript.

teristic distributions, and these distributions differ for surfaces and cores of proteins [7, 66, 69, 76]. In particular, hydrophobic residues tend to be more frequent in the core and polar residues tend to be more frequent on the surface. Further, sites in the core of a protein tend to be more conserved and to evolve slower than surface sites [12, 18, 31, 38, 62, 65, 82, 85]. Presumably, sites in the core tend to be conserved because mutations at these sites are more likely to destabilize the protein fold, due to steric clashes [17].

However, protein properties also vary systematically with factors related to the cellular environment in which proteins are expressed. For example, more highly expressed proteins tend to be more soluble and have less-sticky surfaces [54, 101]. Current protein design algorithms optimize primarily for fold stability [50]. Therefore, we would not expect them to reproduce any patterns caused by the cellular expression environment. By contrast, any patterns that are driven primarily by the requirement for sufficient fold stability, such as avoidance of steric clashes in the core, should be reproduced in computationally designed proteins.

Here, we carried out a systematic comparison between alignments of natural sequences and the corresponding alignments of designed sequences, for several different design conditions. We considered two distinct data sets, one of whole protein structures and one of individual protein domains. We analyzed which design conditions produced sequence alignments that were most similar to natural sequence alignments. We also analyzed by which parameters designed proteins differed the most from natural sequences. Overall, we

found that proteins designed with a flexible backbone and using an intermediate amount of backbone flexibility were the most similar to natural proteins. However, substantial differences between designed and natural proteins remained even under the most advantageous design conditions. In particular, designed proteins tended to have too many polar and too few hydrophobic residues in the core, and they also tended to have cores that were too variable and/or surfaces that were too conserved. These trends were exacerbated for longer proteins.

## 2.2 Materials and Methods

### 2.2.1 Data sets

We analyzed two data sets, one of whole yeast proteins and one of protein domains. The yeast-proteins data set was taken from [82] and comprised 38 protein structures homologous to an open reading frame in *Saccharomyces cerevisiae*. For each of those structures, we had at least 50 homologous natural sequences, also taken from [82]. The protein-domain data set was taken from [68] and comprised 40 protein domains. Only domains with at least one crystal structure in the Protein Database (PDB) and at least 500 sequences in the Pfam Database were selected for this data set. Domains were selected in order to represent several different types of protein folds and domains were also restricted to a length less than or equal to 150 amino acids. For each of these protein domains, we obtained alignments of homologous natural sequences from the Pfam database (Pfam), as described [68].

### 2.2.2 Protein design

For each structure in both data sets, we computationally designed 500 variants each, using multiple design methods. All design methods we used are implemented in the protein-design software Rosetta [53]. First, we used standard fixed-backbone design [51]. In this method, the protein backbone remains fixed and only amino-acid side chains are allowed to move. Second, we used the flexible-backbone method Backrub [93], which first generates an ensemble of alternative backbones and then designs side chains onto these backbones [34, 94]. The Backrub method takes as input a temperature parameter that determines the extent of backbone movements that occur during design. Here, we used temperatures spanning from 0.03 to 2.4 corresponding to increasing backbone movements. For the protein-domain data set, we also used one additional design method, called “Soft”. This method keeps the backbone fixed but the energy function used during sequence design dampens the weight of the repulsive Lennard-Jones (LJ) potential term [68]. Protein designs for the protein-domain data set have been previously published [68], while the designs for the yeast-proteins data set were newly generated for the present study.

All designs for the yeast-proteins data set were generated with Rosetta Revision 39284. For fixed-backbone design, we used the following command:

```
./fixbb.linuxgccrelease -database rosetta_database \  
-s input.pdb -resfile ALLAA.res -ex1 -ex2 \
```

```
-extrachi_cutoff 0 -nstruct 1 -linmem_ig 10
```

Flexible-backbone design was performed by generating a conformational ensemble of 500 structures and then using fixed-backbone design to predict a low energy sequence for each structure in the ensemble. To generate structures for the conformational ensemble, we used the following command:

```
./backrub.linuxgccrelease -database rosetta_database \
    -s input.pdb -resfile NATAA.res -ex1 -ex2 \
    -extrachi_cutoff 0 -backrub:mc_kt <T> \
    -backrub:ntrials 10000 -nstruct 1 -backrub:initial_pack
```

where <T> has to be replaced by the desired design temperature.

The design details for the protein-domain data set can be found in [68].

### 2.2.3 Data analysis

We quantified the variability of sites in amino-acid alignments using site entropy  $H_i$ , defined as  $H_i = -\sum_j p_{ij} \ln p_{ij}$ . Here,  $p_{ij}$  is frequency of amino acid  $j$  in alignment column  $i$ , and the sum runs over all amino acids. We compared amino-acid distributions of designed sequences to those of natural sequences using the Kullback-Leibler (KL) divergence. The KL divergence  $D_i^{\text{KL}}$  is defined as  $D_i^{\text{KL}} = -\sum_j p_{ij} \ln(p_{ij}/q_{ij})$ , where  $q_{ij}$  is the frequency of amino acid  $j$  in column  $i$  of the reference alignment, and  $p_{ij}$  is the corresponding frequency in the alignment that is being compared to the reference alignment. The sum



runs over all amino acids. When calculating frequencies used for the KL divergence we corrected for the presence of frequencies of zero by adding  $1/20$  to each amino acid count before calculating the frequencies. The KL divergence is inherently an asymmetric distance measure, comparing a probability distribution of interest to a reference distribution. Unless noted otherwise, we always used natural sequence alignments to calculate the reference frequencies  $q_{ij}$  and designed sequence alignments to calculate the frequencies  $p_{ij}$ . Throughout this work, we calculated  $D_i^{\text{KL}}$  separately at each site  $i$  in a protein, and then averaged the  $D_i^{\text{KL}}$  values for all sites in a protein to obtain a mean KL divergence for that protein.

To compare the shapes of amino-acid distributions while disregarding specific amino-acid identities, we performed a second type of KL calculation where we ordered amino-acids by their relative frequencies. Thus, instead of the frequencies  $p_{ij}$  and  $q_{ij}$  we used  $p_{ir_j}$  and  $q_{is_j}$ , where  $r_j$  is the rank of the frequency of amino acid  $j$  in column  $i$  of the alignment being compared to the reference, and  $s_j$  is the rank of the frequency of amino acid  $j$  in column  $i$  of the reference alignment. This way of calculating the KL divergence compares the frequencies of amino acids at equal frequency rank, regardless of which specific amino acids are the most frequent, second-most frequent, and so on in each alignment. As an example, assume that at a given site there are only three different amino acids in the natural alignment, I, L, and V, at frequencies 0.5, 0.35, and 0.15, respectively. At the same site in the designed sequences, there are amino acids A, V, and I, also at frequencies 0.5, 0.35, and 0.15, respectively.

In our calculation of KL divergence comparing amino acids at equal frequency rank, we would then compare the frequency of I in the natural alignment with the frequency of A in the designed alignment (the two most frequent amino acids in the two respective alignments) and similarly the frequency of L with the frequency of V and the frequency of V with the frequency of I, respectively. In this example, since the two sets of three frequencies are exactly the same if we disregard amino-acid type, we would obtain a KL divergence of zero.

We calculated Relative Solvent Accessibility (RSA) of residues by first calculating the absolute Solvent Accessibility (ASA) for each residue, using the software DSSP [46]. For each protein, we extracted the chain of interest from the PDB structure and ran DSSP only on that chain. We calculated RSA by dividing the ASA value for each residue by the maximum possible ASA value, as given by [103]. Throughout this work, we only calculated RSA on the native PDB structure. We did not perform any RSA calculations on designed structures. All our data and analysis scripts are available online at: [https://github.com/clauswilke/protein\\_design\\_and\\_site\\_variability](https://github.com/clauswilke/protein_design_and_site_variability).

## 2.3 Results

We wanted to assess the extent to which the sequence space of computationally designed proteins overlaps with the sequence space occupied by homologous natural proteins. Our general approach was to compare alignments of designed protein sequences to alignments of homologous natural sequences, for approximately 80 distinct protein structures. For each structure,

we considered several different design methods (see Methods for details), and we designed 500 sequences for each structure and method. The protein structures we considered were subdivided into two distinct data sets, a data set of 38 yeast protein structures previously analyzed by [82] and a data set of 40 protein domains previously analyzed by [68]. Throughout this study, we analyzed these two data sets separately, because they corresponded to structures of substantially different sizes. The mean number of amino acids per structure was 215.4 in the yeast-proteins data set and 86.1 in the protein-domains data set. Also, the overall sequence variability of the protein-domain data set was greater than the variability of the yeast-proteins data set.

### **2.3.1 Overall site variability**

We first compared overall amino-acid variability in designed and natural proteins. We assessed amino-acid variability at individual sites by calculating the entropy  $H_i$  at each site  $i$  in alignments of either designed or natural proteins. We then calculated the mean entropy over all sites in each alignment and used that quantity as a measure of the overall amino-acid variability in the alignment.

We found that protein design using a fixed backbone generally yielded insufficient site variability compared to natural sequences (Figure 2.1). This result was magnified in the smaller protein domains. In fact, for the protein domains, the most variable proteins under fixed-backbone design showed only about as much variability as the least variable natural proteins. Overall, there

was a significant shift towards higher variability in natural proteins relative to proteins designed with fixed backbone (paired  $t$  test,  $P = 1.4 \times 10^{-10}$  for the yeast-proteins data set and  $P < 10^{-15}$  for the protein-domain data set). When switching from fixed-backbone design to variable-backbone design, we found that overall site variability increased. Further, site variability increased monotonously with the degree of backbone flexibility allowed during design, as measured by the design temperature (Figure 2.1). At the highest temperatures, site variability in designed proteins consistently exceeded that of natural proteins.

Proteins designed at intermediate temperatures had site variabilities that most closely resembled that of natural proteins. For the yeast-proteins data set, the temperature that provided the closest match was  $T = 0.03$ , even though the variability of sequences designed at that temperature still exceeded the variability in natural sequences (paired  $t$  test,  $P = 0.0006$ ). For the protein-domains data set, the temperature that provided the closest match was  $T = 0.9$ , for which variability was statistically indistinguishable from that found in natural sequence alignments (paired  $t$  test,  $P = 0.353$ ). However, for both data sets, natural sequences generally showed a larger spread in variabilities than did the designed sequences at the closest-matching temperatures (Brown-Forsythe test for equal variances,  $P = 0.0003$  for the yeast-proteins data set at  $T = 0.03$  and  $P = 7.3 \times 10^{-6}$  for the protein-domain data set at  $T = 0.9$ ).

### 2.3.2 Amino-acid distributions

We next compared amino-acid distributions between designed and natural sequences. First we looked at overall amino acid frequencies. We found that by-and-large, amino acid frequencies in designed proteins mirrored those in natural proteins (Figure 2.2 and Figures 2.3–2.7). The biggest differences arose in Pro, His, Trp, Phe, and Ala. (We ignore Cys here because Cys is never used in the design algorithm and thus always at frequency 0.) Overall, we observed that hydrophobic residues tended to be under-represented in designed proteins whereas hydrophilic residues tended to be over-represented. This trend was stronger in the protein core than on the surface. We also observed that the longer proteins in the yeast-proteins data set showed larger deviations between designed and natural sequences than the shorter proteins in the protein-domains data set. Finally, when comparing different design methods and design temperatures, we found that differences in amino-acid distributions were relatively minor, see Figure 2.2 and Figures 2.3–2.7.

Even if overall amino-acid distributions are approximately correct, the amino-acid distributions at individual sites can be poorly predicted [82]. Therefore, we next compared, separately at each site, the similarity between amino-acid distributions in natural proteins and those in designed proteins. To carry out this comparison, we employed the Kullback-Leibler (KL) divergence [108], which measures how similar one probability distribution is to a reference distribution. A KL divergence of zero implies that the distributions are identical. The higher the KL divergence, the more dissimilar the focal distribution is to

the reference distribution. (Note that KL divergence is not symmetric: if we swap the focal and the reference distribution, we will generally obtain a different KL divergence value.) We calculated the KL divergence at each site in each protein, and then averaged over sites within a protein to obtain a mean similarity score for each protein. As a control, we also randomly split the alignment of natural sequences for each protein structure into two halves and calculated the mean KL divergence of natural sequences against themselves.

First, in all comparisons, we found that the KL divergence of designed relative to natural sequences was much bigger than the KL divergence of natural sequences relative to themselves (Figures 2.8 and 2.9). This finding indicates a substantial discrepancy between designed and natural sequences at individual sites. Second, we found that the mean KL divergence decreased with increasing design temperature (Figures 2.8A and 2.9A). Thus, according to the KL divergence measure, structures designed with the most flexible backbones had the most similar amino-acid distributions to those found in natural sequences.

However, the result that sequences designed at the highest temperatures are the most similar to natural sequences may be an artifact of the KL divergence measure. As design temperature increases, amino-acid variability increases, and amino-acid distributions become more uniform. A more uniform distribution is generally going to display more overlap with any given distribution than a more localized distribution, if the localized distribution is not correct. Thus, the decrease in KL divergence with increasing temperature may

simply reflect the broadening of the distribution, not an actual improvement in reproducing natural amino-acid distributions. To assess whether amino-acid distributions in designed sequences were simply broadening with increasing temperature, or whether they were actually converging on the natural distributions, we carried out a second set of comparisons. We rank-ordered amino acids by frequency at each site in each protein, and then calculated the KL divergence of the rank-ordered distributions.

This comparison considers only the shape of the distribution and does not assess whether the correct amino acids are present at individual sites. This second comparison generally found much lower KL divergence levels, even though still not as low as what was found for the control comparison of natural sequences with themselves (Figures 2.8B and 2.9B). More importantly, now KL divergence reached a minimum around a temperature of 0.3 (yeast proteins, Figure 2.8B) to 1.2 (protein domains, Figure 2.9B) and rose again beyond that value. This finding indicates that higher design temperatures do not unequivocally produce more natural amino-acid distributions. Instead, there is an intermediate temperature, approximately coinciding with the temperature at which overall sequence variability matches best, at which amino acid distributions also are most similar.

### **2.3.3 Site variability and solvent accessibility**

The previous analyses demonstrated that while designed proteins overall look similar to natural proteins, there are also important differences. We

next wanted to identify whether these differences were present uniformly throughout the structure or could be located to specific structural regions. In our analysis of amino-acid distributions, we had already seen that amino-acid distributions seemed to deviate more at buried sites than at exposed sites (Figures 2.2 and 2.6).

We first plotted site variability against relative solvent accessibility (RSA, a dimensionless number from 0 to 1 measuring the relative solvent exposure of individual residues) for individual proteins. See Figure 2.10 for one example. We generally found that site variability displayed a substantial spread even for sites of very similar RSA. At the same time, there was an overall trend for sites with higher RSA to be more variable than sites with lower RSA. This trend was generally stronger in flexible backbone designs than in fixed backbone designs (Figure 2.10). To analyze the relationship between site variability and RSA more systematically, we calculated the correlation between these two quantities for all proteins (Figure 2.11 and 2.12). On average, natural sequence alignments showed a higher correlation than alignments of designed sequences, regardless of design method.

Intermediate design temperatures showed the highest correlations, but correlations were nevertheless significantly lower in designed proteins than in natural proteins (paired  $t$  test,  $P = 2.96 \times 10^{-10}$  [ $T = 0.3$ , yeast proteins] and  $P = 1.75 \times 10^{-5}$  [ $T = 0.3$ , protein domains]). We also investigated whether the designed proteins with the highest correlations corresponded to the natural proteins with the highest correlations, and found this generally to be the case



(Figures 2.11B and 2.12B).

Our finding that correlations between site entropy and RSA are lower in designed proteins than in natural proteins indicates that, in designed proteins, site variability is too uniform across different solvent exposure states. In short, designed proteins are either too variable in the core or too conserved on the surface. To obtain a clearer picture of how exactly designed proteins differed from natural proteins, we once more considered the distributions of mean site entropies, but now calculated separately for buried sites ( $\text{RSA} \leq 0.05$ ), for partially buried sites ( $0.05 < \text{RSA} \leq 0.25$ ), and for exposed sites ( $\text{RSA} > 0.25$ ). Figure 2.13 shows the medians of these distributions. For designed proteins, the mean site variabilities of exposed and of partially buried sites are close in magnitude while the mean site variabilities of buried sites are generally consistently lower. By contrast, in natural sequences exposed sites show much more variability than partially buried sites.

If buried sites are too variable or exposed sites too conserved in designed proteins, we reasoned that hybrid designs, in which buried sites were taken from sequences designed at a lower temperature and exposed sites from sequences designed at a higher temperature, should display correlations more similar to those seen in natural proteins.

According to Figure 2.13, for the yeast proteins buried and partially buried sites in designed proteins had site variability most similar to that of natural sequences in proteins designed with a fixed backbone or in proteins with a design temperature of  $T = 0.03$ . In the protein-domains data set,

that temperature was  $T = 0.3$  to  $T = 0.6$ . By contrast, for exposed sites the site variability in designed proteins was most similar to that of natural sequences at a design temperature of  $T = 0.1$  (yeast proteins) and  $T = 1.2$  (protein domains). We thus built our hybrid designs by combining sites from these temperatures. We found that the distribution of the site-entropy–RSA correlations in hybrid designs was comparable to that in natural sequences (Figure 2.14). However, predictions for specific proteins lacked accuracy (Figure 2.15).

## 2.4 Discussion

We have compared site variability and amino-acid distributions in designed and natural proteins, for two distinct data sets. One data set consisted of 38 yeast proteins and the other consisted of 40 protein domains. Structures in the yeast-proteins data set were, on average, much larger than structures in the protein-domain data set, while natural sequences in the protein-domain data set were more variable than those in the yeast-proteins data set. We have found that proteins designed with a flexible backbone, using an intermediate design temperature, were generally the most similar to natural proteins. Overall amino-acid frequencies in designed proteins were similar, though not identical, to those in natural proteins. However, amino-acid frequencies at individual sites showed substantial deviations. Finally, we have found that site variabilities in designed proteins are too uniform across different solvent exposure states of residues. Designed proteins have either cores that are too

variable or surfaces that are too conserved.

In previous studies, native sequence recovery has been used to assess design accuracy [35, 50]. Native sequence recovery is defined as the mean percent of native amino acid identities that are observed in the designed proteins. Despite its widespread use, native sequence recovery may not always be a sufficient indicator of design accuracy, especially when examining different sequences that are compatible with one specific structure. A major goal of design is to find sequences that fold into a specific structure. For this goal, one typically models a series of structures that are similar to the native structure and then identifies low energy sequences for each of these modeled structures. Even if all designed sequences fold into the desired structure, they may not necessarily have a high sequence similarity with the sequence of the native structure. For this reason, we believe that it is important to assess design accuracy by multiple different methods, and also against an ensemble of native sequences or structures.

A previous study, the source of the protein-domains data set we analyzed here, has similarly compared designed proteins against ensembles of natural sequences [68]. That study and our present study complement each other. [68] were primarily interested in amino-acid covariation, and they also considered sequence entropy and profile similarity [113]. Here, we were primarily interested in the effects of solvent occlusion on site variability and amino-acid choice, and we also considered two distinct sets of natural reference structures (protein domains and whole proteins). In both studies, an

intermediate amount of backbone flexibility was found to be optimal for recapitulating characteristics of natural protein sequences. Both studies also identify similar inaccuracies in the designed protein sequences. [68] observed that covarying pairs in designed protein cores were more likely to be hydrogen bonding pairs than in natural cores, and here we found that polar residues are over-represented in the designed protein cores compared to natural cores.

Our analysis compared two distinct datasets. The first was comprised of 40 protein domains, chosen to be less than 150 amino acids in length and with a mean length of 86.1 amino acids. The second was comprised of 38 whole yeast proteins, with a mean length of 215.4 amino acids. For each structure in each data set, we had an associated alignment of natural sequences to assess natural variability for that structure. (Note that sequences homologous to the yeast proteins were not constrained to be fungal sequences.) Sequences in the protein-domain data set were more variable than sequences in the yeast-protein data set. We found that optimal design temperatures were lower for the yeast-protein data set than for the protein-domains data set. This finding is consistent with both increased mean length and reduced mean variability in the yeast-protein data set relative to the protein-domains data set. In particular, large cores in the larger proteins may lead to larger conserved regions whose site variability patterns are better recaptured at lower design temperatures.

We found that the characteristics of designed protein sequences are generally similar but by no means identical to natural sequences. To some extent, this discrepancy is to be expected. Designed protein sequences are op-

timized entirely for thermodynamic stability as estimated by the design energy function. Natural proteins experience a variety of selective pressures, stability being only one of them. For example, natural proteins experience selection pressures for native protein–protein interactions, against non-specific protein–protein interactions, and against misfolding and aggregation [23, 33, 54, 114]. If they are enzymes, natural proteins also require the appropriate mutations that enable enzymatic activity, even if those mutations are thermodynamically destabilizing [8, 26]. While selection for enzymatic activity will likely affect only a few sites in a protein, the other selective forces (misfolding, aggregation, native and non-specific interactions) have the potential to exert much broader selection pressures across many sites in a protein. As long as design algorithms do not take these selection pressures into account, we cannot expect design algorithms to reproduce natural sequence variation exactly.

To identify at what sites discrepancies between natural and designed proteins arose, we explicitly examined the relationship between structure and sequence variability. In particular, we analyzed the correlation between RSA and site entropy, which reflects the well-known observation that proteins are more variable on the surface than in the core. We found that the difference between surface and core variability was much more pronounced in natural proteins than in designed proteins. Designed proteins either have cores that are too variable or surfaces that are too conserved. We created hybrid designs, taking core sites from one set of designed proteins and surface site from another set, designed with more backbone movement, and tested whether these

hybrid designs showed the appropriate differential in variability between core and surface sites. We found that they did so as a population (Figure 2.14) but not individually (Figure 2.15). This observation indicates that there is some aspect of protein fold stability that differentially affects surface and core residues and that is not yet properly incorporated into current design algorithms. Simply raising the design temperature on the surface but not in the core is not sufficient to capture this effect. Note that we do not expect our hybrid design approach to yield realistic, stable protein sequences. It is merely meant as an illustration of the extent to which surface sites would have to be more variable relative to core sites to yield entropy-RSA correlations comparable to those found in natural sequences.

For both data sets, the designed proteins had fewer hydrophobic residues and more polar residues than expected from natural sequence alignments. This trend was particularly apparent in the protein core, and it was more extreme for larger proteins. These discrepancies suggest a need for further improvement of the design algorithm, most likely the energy function. Rosetta uses a scoring function that predicts the energy of a given sequence folded into a particular target structure [51]. As a component of this scoring function, Rosetta uses the Lazaridis-Karplus implicit solvation model to estimate the energy of desolvation of each residue [52]. The over-representation of polar residues in protein cores that we observed suggests that this solvation model is either insufficiently penalizing for the burial of polar groups or insufficiently rewarding the burial of hydrophobic residues. Improvements to the solvation model

used in design may result in more stable designed proteins with amino acid distributions more similar to those of natural proteins, especially in protein cores.

While protein cores are more variable in designed proteins compared to natural proteins, the surfaces of designed proteins are too conserved. This discrepancy is somewhat expected. We would only expect close agreement between designed and natural proteins if the sequences are under the same constraints (and provided the energy function could accurately capture these). Computational design optimizes sequences primarily for protein stability, which, in natural proteins, is more likely to be a dominant constraint in protein cores than on surfaces. Surfaces of natural proteins may also be under other important pressures, such as to make desired and avoid unwanted interactions and to keep proteins soluble. All of these pressures could act to diversify protein surfaces away from sequence choices that would maximize stability. In addition, there are of course also inaccuracies in the design energy function, including difficulties in accurately modeling electrostatics and solvation at surfaces, and contributions of conformational entropy of surface side chains that are not taken into account in most design energy functions.

In our analysis of approximately 80 protein structures total, we found that proteins designed with an intermediate amount of backbone flexibility exhibited site-variability patterns most closely resembling that of natural proteins. However, the optimal range of backbone flexibility was different in the two data sets. Further, even when the overall site variability matched that of

natural sequences, the specific amino-acid distributions at individual sites did not match that well, as quantified by the relatively large KL divergence values between natural and designed alignments. Similarly, intermediate design temperatures showed the highest correlation between RSA and site variability (as measured by entropy). However, even at the optimal design temperature ( $T \sim 0.3$  for both data sets), the designed proteins exhibited systematically lower correlations than did the natural proteins. Consequently, using current state-of-the-art design algorithms, designed proteins have either surfaces that are too conserved or cores that too variable. We suspect that changes in the design energy function, in particular more accurate estimation of the balance between electrostatics and desolation, will be needed to address this issue. We also see a need for improved flexible-backbone design algorithms that can model larger backbone movements on the surface without disturbing the core backbone as much. As alternative and improved algorithms design algorithms become available, they should be subjected to similar tests as we have done here, to assess to what extent different algorithms reproduce natural amino-acid frequency and site-variability differences in core versus surface.



## 2.5 Figures

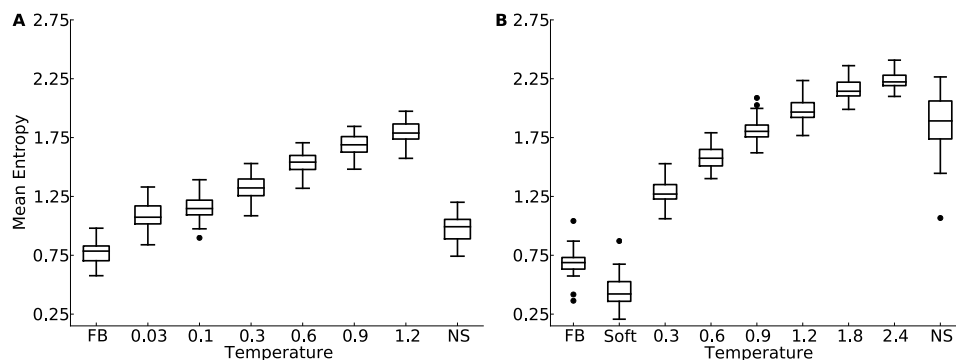


Figure 2.1: Mean site entropy for designed and natural proteins. Each boxplot represents the distribution of mean site entropies within the respective dataset (A: yeast proteins; B: protein domains). “FB” refers to fixed-backbone design. Temperature values refer to the design temperature used during the Backrub design method. “NS” refers to natural sequences. “Soft” refers to the Soft design method. We find generally that increased backbone flexibility allows for more site variability. Intermediate temperatures produce site variabilities most similar to those seen in natural sequences. Overall, natural sequences in the protein-domains data set are more variable than are those in the yeast-proteins data set.

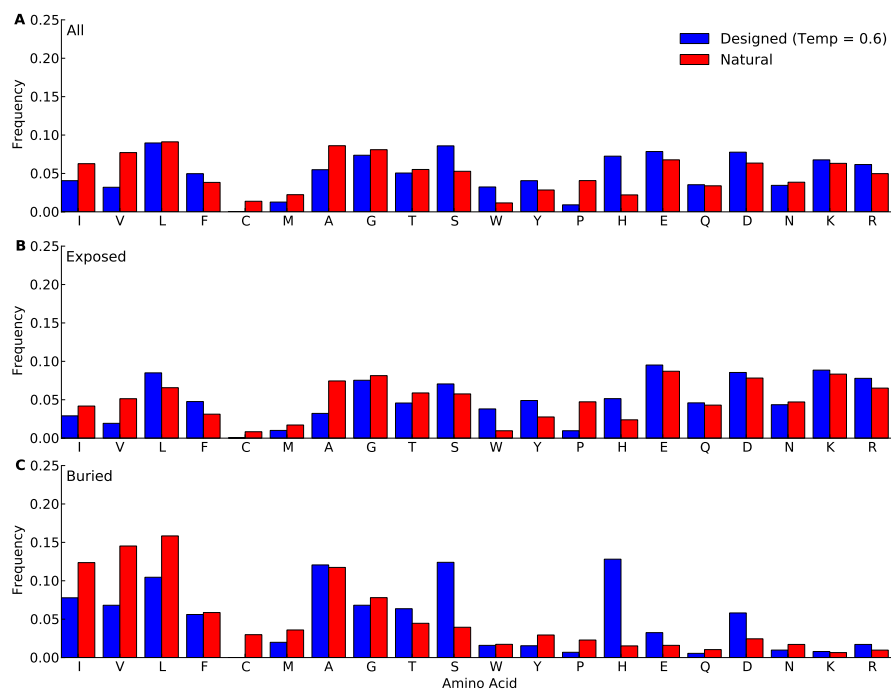


Figure 2.2: Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the yeast-proteins data set. For designed proteins, only flexible-backbone designs with design temperature 0.6 were considered. (A) Overall frequencies. (B) frequencies at exposed sites (defined as sites with RSA > 0.05). (C) frequencies at buried sites (defined as sites with RSA ≤ 0.05).

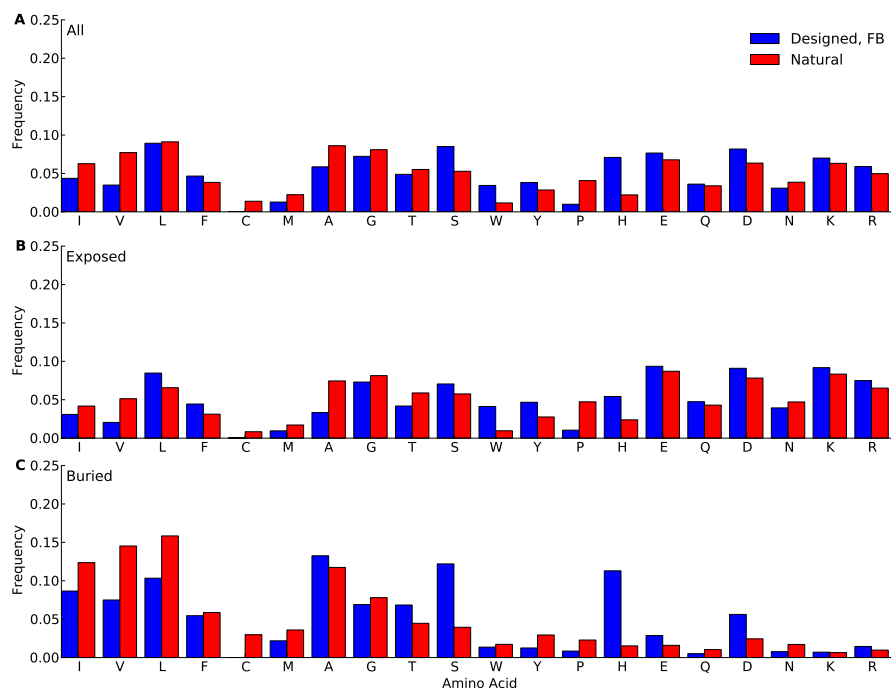


Figure 2.3: Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the yeast-proteins data set. For designed proteins, only fixed-backbone designs were considered. (A) overall frequencies. (B) frequencies at exposed sites (defined as sites with  $\text{RSA} > 0.05$ ). (C) frequencies at buried sites (defined as sites with  $\text{RSA} \leq 0.05$ )

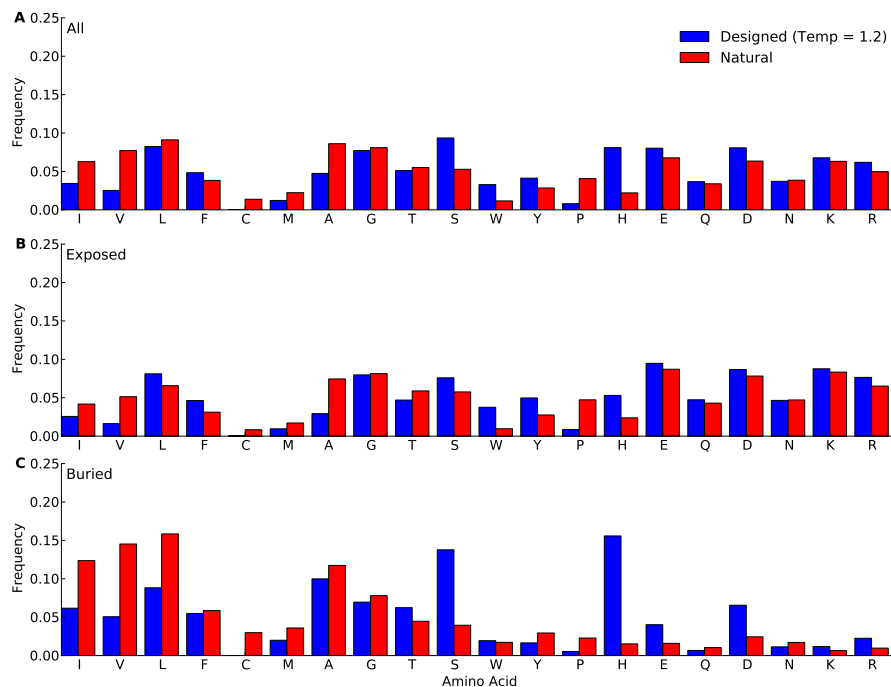


Figure 2.4: Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the yeast-proteins data set. For designed proteins, only flexible-backbone designs with design temperature 1.2 were considered. (A) overall frequencies. (B) frequencies at exposed sites (defined as sites with  $\text{RSA} > 0.05$ ). (C) frequencies at buried sites (defined as sites with  $\text{RSA} \leq 0.05$ ).

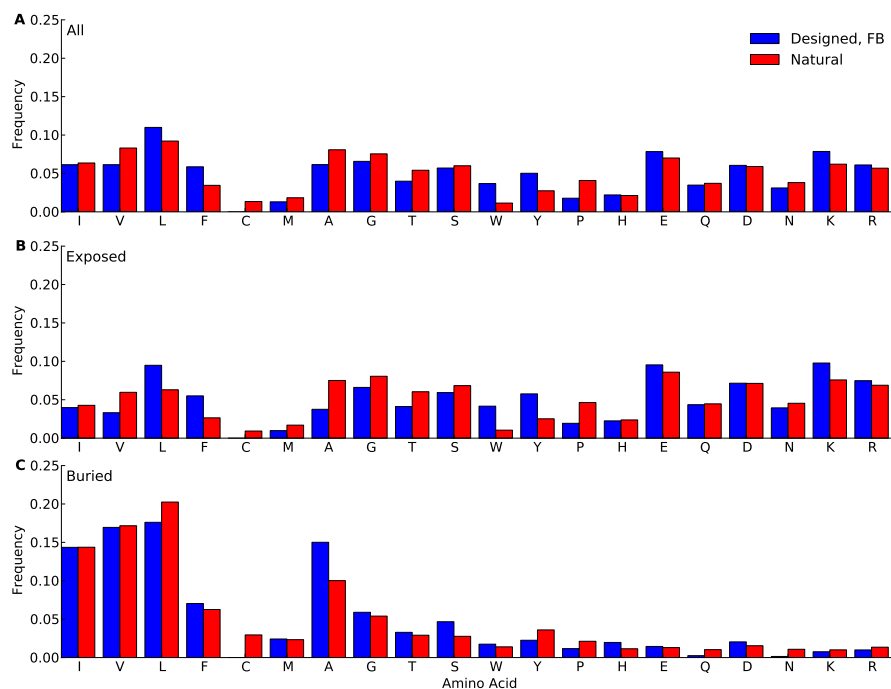


Figure 2.5: Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the protein-domains data set. For designed proteins, only fixed-backbone designs were considered. (A) overall frequencies. (B) frequencies at exposed sites (defined as sites with  $\text{RSA} > 0.05$ ). (C) frequencies at buried sites (defined as sites with  $\text{RSA} \leq 0.05$ )

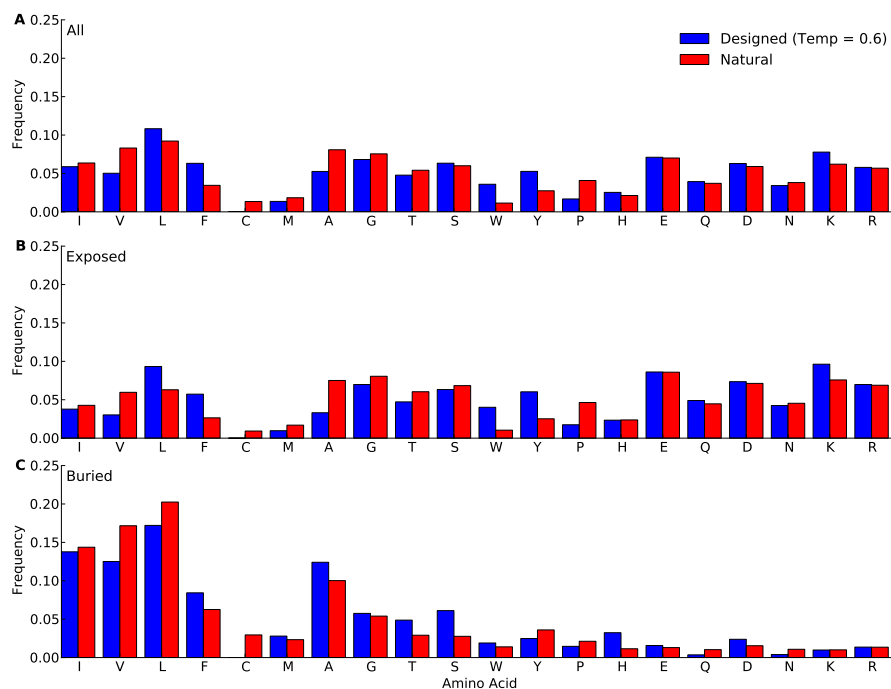


Figure 2.6: Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the protein-domains data set. For designed proteins, only flexible-backbone designs with design temperature 0.6 were considered. (A) overall frequencies. (B) frequencies at exposed sites (defined as sites with RSA > 0.05). (C) frequencies at buried sites (defined as sites with RSA ≤ 0.05)

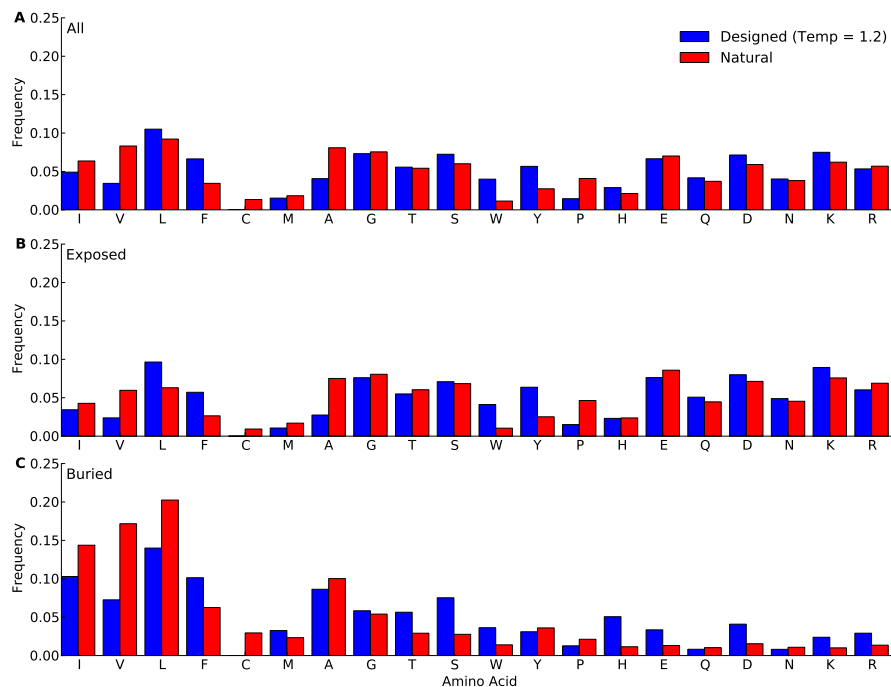


Figure 2.7: Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the protein-domains data set. For designed proteins, only flexible-backbone designs with design temperature 1.2 were considered. (A) overall frequencies. (B) frequencies at exposed sites (defined as sites with RSA > 0.05). (C) frequencies at buried sites (defined as sites with RSA ≤ 0.05)

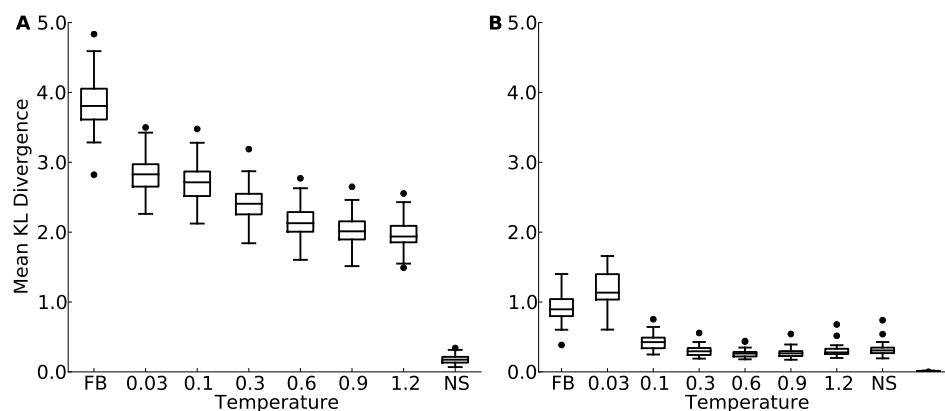


Figure 2.8: Mean Kullback-Leibler (KL) divergence for designed and natural proteins, shown for the protein-domain data set. A higher KL divergence indicates that the amino-acid distributions at sites in designed proteins are less similar to the corresponding distributions in the natural proteins. “FB” refers to fixed backbone design and “NS” refers to the control case where natural sequences are compared to themselves. (A) KL divergence calculated from the relative frequencies of the 20 amino acids. (B) KL divergence calculated from rank-ordered frequency distributions. The most common amino acid in the reference distribution is compared to the most common amino acid in the focal distribution, the same is done for the second-most common amino acid, and so on, irrespective of the type of amino acids.



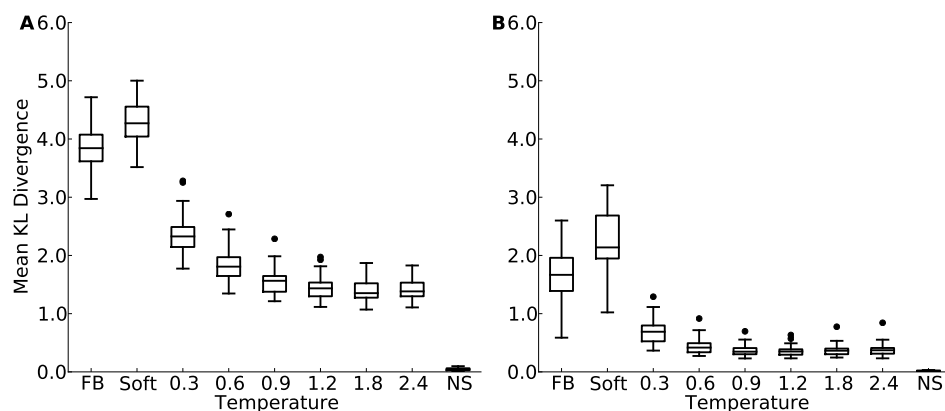


Figure 2.9: Mean Kullback-Leibler (KL) divergence for designed and natural proteins, shown for the yeast-proteins data set. A higher KL divergence indicates that the amino-acid distributions at sites in designed proteins are less similar to the corresponding distributions in the natural proteins. “FB” refers to fixed backbone design, and “NS” refers to the control case where natural sequences are compared to themselves. (A) KL divergence calculated from the relative frequencies of the 20 amino acids. (B) KL divergence calculated from rank-ordered frequency distributions. The most common amino acid in the reference distribution is compared to the most common amino acid in the focal distribution, the same is done for the second-most common amino acid, and so on, irrespective of the type of amino acids.

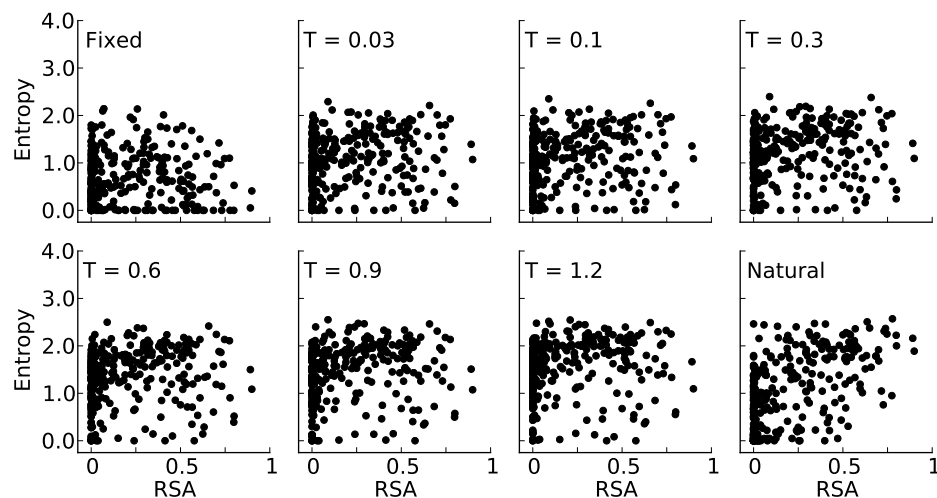


Figure 2.10: Site entropy versus Relative Solvent Accessibility (RSA) for designed and natural sequence alignments of the protein S-formylglutathione hydrolase (PDB: 1PV1, chain A). Natural sequences exhibit a clear trend of higher site variability at higher RSA values. The flexible backbone designs exhibit a similar trend but the fixed backbone designs do not.

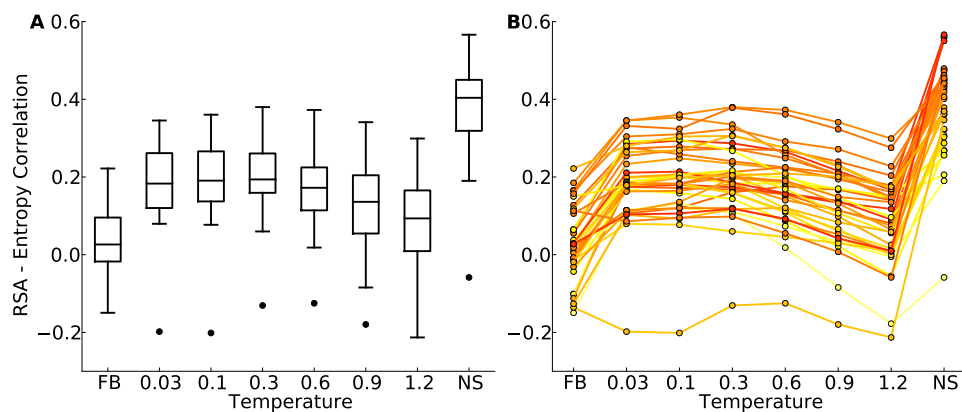


Figure 2.11: Distributions of correlation coefficients between site entropy and RSA, for the protein-domain data set. “FB” indicates fixed-backbone design and “NS” indicates natural sequences. (A) Distributions represented as boxplots. (B) Correlation coefficients for individual proteins. Lines connect identical structures in the different design conditions. The color shading represents the strength of the correlation for the natural sequence alignment. In general, natural proteins display a stronger correlation between site entropy and RSA than designed proteins.

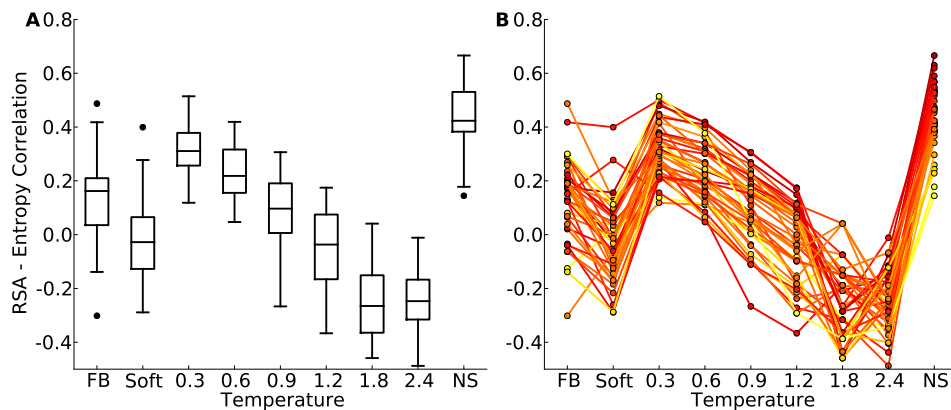


Figure 2.12: Distributions of correlation coefficients between site entropy and RSA, for the yeast-proteins data set. “FB” indicates fixed-backbone design, “Soft” indicates soft backbone design, and “NS” indicates natural sequences. (A) Distributions represented as boxplots. (B) Correlation coefficients for individual proteins. Lines connect identical structures in the different design conditions. The color shading represents the strength of the correlation for the natural sequence alignment. In general, natural proteins display a stronger correlation between site entropy and RSA than designed proteins.

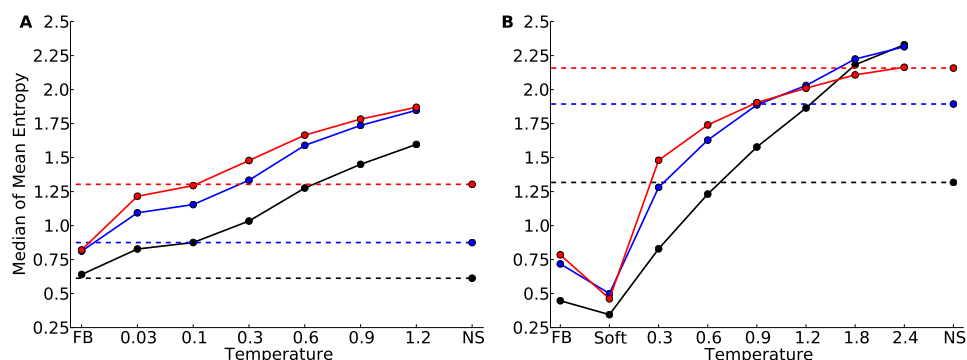


Figure 2.13: Median of the distribution of mean sequence entropies for designed and natural sequences, calculated separately for buried (black), partially buried (blue), and exposed (red) residues (A: yeast proteins; B: protein domains). We defined buried sites as those with  $\text{RSA} \leq 0.05$ , partially buried as those with  $0.05 < \text{RSA} \leq 0.25$ , and exposed as those with  $\text{RSA} > 0.25$ . Dashed lines indicate the corresponding median for natural sequence alignments. Note that for buried (black) and partially buried (blue) residues, the temperatures at which natural site variability and design variability match are comparable. By contrast, for exposed residues, a higher design temperature is required for the design variability to match the natural site variability.

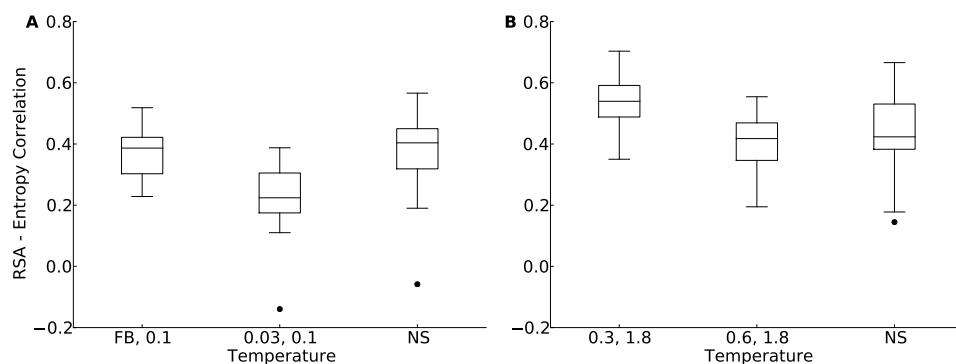


Figure 2.14: Distribution of correlation coefficients between RSA and site entropy for hybrid designs and for natural proteins (A: yeast proteins; B: protein domains). “FB” indicates fixed-backbone design and “NS” indicates natural sequences. For the hybrid designs, buried and partially buried sites were taken from sequences designed at one temperature, and exposed sites were taken from sequences designed at a different temperature. For the hybrid designs, the correlation coefficients were similar to those of natural sequences (paired  $t$  test,  $P = 0.517$  ( $T = \text{FB}, 0.1$ ) and  $P = 6.78 \times 10^{-8}$  ( $T = 0.03, 0.1$ ) [yeast proteins],  $P = 5.19 \times 10^{-5}$  ( $T = 0.3, 1.8$ ) and  $P = 0.118$  ( $T = 0.6, 1.8$ ) [protein domains]).

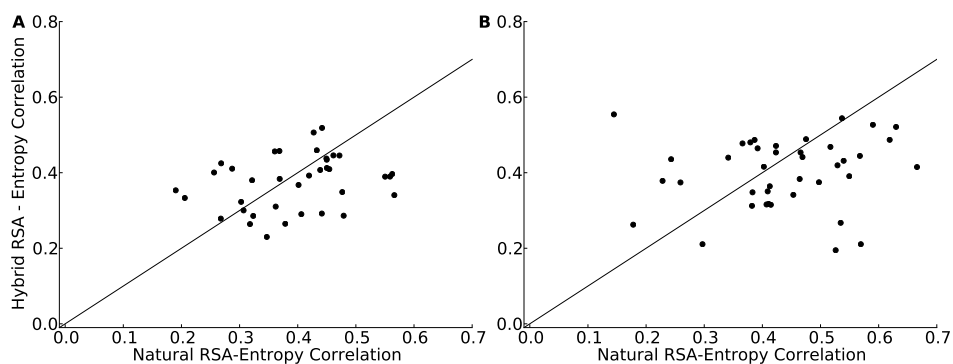


Figure 2.15: Correlation coefficients between RSA and site entropy for hybrid designs and natural proteins. For the hybrid designs, buried and partially buried sites were taken from proteins designed with a fixed backbone (yeast proteins) or a temperature of  $T = 0.6$  (protein domains). Exposed residues were taken from proteins designed with a temperature of  $T = 0.1$  (yeast proteins) or  $T = 1.8$  (protein domains). The solid line indicates  $y = x$ . Note that while the range of correlation values in hybrid designs generally matches the range of values in natural alignments, predictions for specific proteins are not that accurate.

## Chapter 3

# Intermediate divergence levels maximize the strength of structure–sequence correlations in enzymes and viral proteins

### 3.1 Introduction

This work has been previously published in the journal *Protein Science*.<sup>1</sup>

Proteins are subject to a number of biophysical and functional constraints that influence their evolutionary trajectories [56, 90, 109, 115]. These constraints contribute to observed patterns in both whole-gene evolutionary rate variation [8, 23, 33, 55, 87] and evolutionary rate variation among sites within individual proteins [24, 31, 42, 88, 111]. Such evolutionary rate variation in turn contributes to heterogeneity in site-specific sequence variability.

A number of studies have sought to understand the roles that biophysical constraints, particularly structural constraints, play in this observed

---

<sup>1</sup>Eleisha L. Jackson, Amir Shahmoradi, Stephanie J. Spielman, Benjamin R. Jack, and Claus O. Wilke. Intermediate divergence levels maximize the strength of structure–sequence correlations in enzymes and viral proteins. *Protein Science*, DOI: 10.1002/pro.2920, 2016. Amir Shahmoradi helped design the project. Stephanie J. Spielman and Ben R. Jack helped analyze the data and write the manuscript. C. O. Wilke helped to design the project and write the manuscript.



site-specific variability within proteins. Structural properties such as solvent exposure and packing density have emerged as strong predictors of site-wise evolutionary rates [85, 88, 111, 112]. Solvent exposure is typically measured with the metric relative solvent accessibility (RSA), which indicates the extent to which a given residue comes into contact with solvent (i.e., water) [103]. Residues that are exposed on the surface of the protein have high RSA, with complete exposure indicated with an RSA of one. Residues that are buried and/or in the protein core have low RSA, with completely buried residues having an RSA of zero. RSA has a significant, positive relationship with evolutionary rate, such that more buried residues tend to evolve more slowly than exposed residues do [12, 31, 32, 38, 65, 69, 82, 85].

Alternatively, packing density indicates how tightly packed a given residue is by neighboring amino acids in a protein’s tertiary structure. A residue’s packing density is commonly measured as weighted contact number (WCN), which is defined as the sum of the inverse square distance of all residues in the protein to the focal amino acid [57, 89]. Recent work has suggested that WCN is a strong determinant of site-specific variability in proteins, and that residues with high WCN evolve more slowly than do residues with low WCN [42, 55, 111, 112].

However, some studies have yielded apparently contradictory results regarding the extent of the predictive power that these structural properties have on site-wise evolutionary rate (ER). For example, Yeh *et al.* [111] investigated structure–sequence relationships in a data set of 216 monomeric enzymes, find-

ing that WCN is a stronger determinant of site-wise ER than RSA, although RSA was still a significant predictor. Importantly, Yeh *et al.* [111] recovered strong correlations between structure and ER, with WCN and RSA explaining up to  $\sim 41\%$  of the variance in site-specific ER. By contrast, Shahmoradi *et al.* [88] examined the structure–sequence relationship on a set of 9 viral proteins. While Shahmoradi *et al.* [88] similarly found that both RSA and WCN are significant predictors of rate in proteins, the correlations Shahmoradi *et al.* [88] observed were much smaller in magnitude [88]. Specifically, they found that at best, structural predictors could explain only  $\sim 15\%$  of the variance in ER. Given these disparate findings, it remains unclear which of the two studies is the more representative one.

Although both Yeh *et al.* [111] and Shahmoradi *et al.* [88] examined the relationship between sequence and structural properties, they used different methods and data sets. First, Yeh *et al.* [111] measured ER using the method Rate4Site [61, 78], whereas Shahmoradi *et al.* [88] focused on sequence entropy, which is not a rate. Second, Yeh *et al.* [111] used a much more comprehensive data set of monomeric enzymes, and Shahmoradi *et al.* [88] analyzed a comparatively smaller set of viral proteins, which are subject to an additional layer of selective forces imposed by the host immune system. Finally, Shahmoradi *et al.* [88] considered additional structural predictors, namely protein design and flexibility, while Yeh *et al.* [111] focused on RSA and WCN alone. This use of different methods makes it difficult to directly compare results from the two studies.

Here, we attempt to reconcile these two studies, by re-analyzing both the enzyme data set from Yeh *et al.* [111] and the virus data set from Shahmoradi *et al.* [88] in one consistent analysis pipeline. We focus on three structural predictors from the two studies: WCN, RSA, and variability in designed proteins. We confirm that, indeed, correlations between rate and structural predictors are much smaller for the viral proteins compared to the enzymes. However, differences in structural characteristics do not appear to drive the low predictive power in the viral protein data set. Instead, we find that the enzyme and viral protein data sets primarily differ in the extent of sequence variability in the multiple-sequence alignments (MSAs) used to infer evolutionary rates. Using evolutionary models, we quantify sequence divergence for all individual MSAs, and we find that the enzyme data set displays very high levels of divergence while the viral protein data set has experienced minimal evolutionary divergence. Across both data sets, we observe that the strongest structure–sequence correlations are observed at intermediate divergence levels. We conclude that the strength of the structure–structure relationship in proteins is, in part, determined by the extent of sequence variability in the data sets analyzed.

## 3.2 Materials and methods

### 3.2.1 Structures, sequences, and measures of sequence properties

The results presented in this work were based on two data sets. The first was a data set of 208 monomeric enzymes, taken from Echave *et al.* [24] who

re-analyzed the structures originally studied by Yeh *et al.* [111]. The Echave *et al.* [24] data set was slightly smaller than the original data set because Echave *et al.* [24] removed proteins that had missing data at insertion sites. The data set from Echave *et al.* [24] was originally comprised of 209 proteins but we removed one additional protein, 1CQQ, during our analysis (see below for details). Thus, our final enzyme data set had 208 proteins. In brief, these proteins were all enzyme monomers randomly picked from the Catalytic Site Atlas 2.2.11 [75]. Proteins in this data set varied from 95 to 1287 residues in length. Each structure was accompanied by a multiple-sequence alignment of 300 homologous sequences. The second data set was taken from Shahmoradi *et al.* [88] and consisted of nine viral proteins. The viral proteins ranged from 122 to 557 residues in length and each structure was accompanied by a multiple-sequence alignment of up to 2362 homologous sequences. Although both data sets vary in the number of sequence alignments, we did not enforce a medium number sequences in the multiple-sequence alignments needed to be included in the study since all alignments had at least 95 sequences.

Sequence alignments for both data sets were constructed by aligning the amino-acid sequences using the alignment software MAFFT [47,48], specifying the “auto” flag to select the optimal algorithm for the given data set. The alignments were then used to calculate site-specific measures of evolutionary rate for each individual protein in both data sets. We calculated a measure of site-specific evolutionary rate for each protein using the software Rate4Site [61]. First, maximum likelihood phylogenetic trees were inferred with RAxML,

using the LG substitution matrix and the CAT model of rate heterogeneity [98, 99]. For each structure, we used the respective sequence alignment and phylogenetic tree to infer site-specific substitution rates with Rate4Site, using the empirical Bayesian method and the JTT model of sequence evolution [61].

Using the alignments, we also calculated the Shannon entropy ( $H_i$ ), at each alignment column  $i$ :

$$H_i = - \sum_j P_{ij} \ln P_{ij}, \quad (3.1)$$

where  $P_{ij}$  was the relative frequency of amino acid  $j$  at position  $i$  in the alignment. Sequence entropy is a measure of variability at each site.

Finally, we calculated the divergence of each multiple-sequence alignment, using two measures: mean root-to-tip distance and mean patristic distance. Mean root-to-tip distance counts the average number of substitutions that have occurred along the tree. The mean patristic distance of an alignment was the average patristic distance of a tree where patristic distance was defined as the sum of the branch lengths between two nodes (i.e., sequences) within the tree [30]. Both root-to-tip distance and patristic distance were calculated using DendroPy [100].

For the viral proteins we collected a second data set. Using the sequences from the nine viral proteins from Shahmoradi et al. [88] as queries, we used PSI-BLAST [3] against the Uniprot90 to obtain homologous sequences for each protein. We used MAFFT and RAxML to create alignments and

build trees for each protein. Trees could not be created for three of the proteins because their alignments did not have a sufficient number of sequences. We also chose to discard proteins from the analysis that did not have at least 25 sequences. This was done to guard against inaccurate rates. We calculated evolutionary rates for the remaining three proteins (PDB IDs: 1RD8, 3GOL, and 3LYF) using Rate4Site.

We quantified MSA reliability using a re-implementation of the Guidance platform [73] introduced by Spielman *et al.* [95]. Guidance quantifies how robust MSA columns are to the guide tree topology used in during a progressive alignment algorithm. For each MSA column, Guidance produces a column score ranging from 0, indicating that the column is highly unreliable, to 1, indicating that the column is highly reliable. Note that the implementation in Spielman *et al.* [95] differs from that in Penn *et al.* [73] through its use of FastTree [77] to construct perturbed guidetrees. Here, Guidance was run with 100 bootstrap replicates using the MAFFT [47, 48] alignment software, specifying the “auto” flag. We derived an overall Guidance score for each MSA by averaging its resulting Guidance column scores.

### 3.2.2 Protein Design

Using Rosetta [53], we computationally designed 500 structures for select proteins in each data set. For the viral proteins we designed 500 structures for each of the proteins taken from Shahmoradi *et al.* [88]. For the enzymes we designed structures for each protein that was at most 200 residues in length.

For each protein, we first designed 500 flexible ensembles using Backrub [93]. Backrub generates a set of flexible backbone “ensembles” onto which side-chains can then be designed [93,94]. The Backrub method takes a temperature parameter,  $T$ , that determines the extent of backbone flexibility during design. Higher temperatures allow for more backbone flexibility. Previous work has shown that moderate temperature parameters result in designed structures more similar to natural proteins [43,68]. Therefore, we used 0.6 as our temperature parameter. We then used the fixed-backbone method [51] to design side-chains on these ensembles.

All designs were generated with Rosetta 3.5, 2014 week five release. To generate the series of ensembles using flexible-backbone design we used the following Rosetta commands:

```
./backrub -database rosetta_database \
-s input.pdb -resfile NATAA.res -ex1 -ex2 \
  -extrachi_cutoff 0 -backrub:mc_kt 0.6 \
  -backrub:ntrials 10000 -nstruct 1 -backrub:initial_pack
```

For the fixed-backbone design we used the following Rosetta commands:

```
./fixbb -database rosetta_database \
-s input.pdb -resfile ALLAA.res -ex1 -ex2 \
  -extrachi_cutoff 0 -nstruct 1 -overwrite \
  -minimize_sidechains -linmem_ig 10
```

After design, we removed proteins that did not map back properly to the alignments. This resulted in the removal of one structure, 1CQQ, completely from the study. This resulted in a total of 32 enzymes in addition to the viral proteins.

Using the sequence alignments of designed proteins we predicted a site-wise rate, using the expression for dN proposed by Spielman and Wilke<sup>32</sup> [97] (as implemented in the software Pyvolve [96]). For this calculation, we assumed that the mutation rate at all sites was equal. We called this quantity the “design rate” (DR) at sites.

### 3.2.3 Calculation of structural properties

In our analysis, we used side-chain Weighted Contact Number (WCN) as proposed by Marcos and Echave [59]. This quantity is defined as

$$\text{WCN}_i = \sum_{i \neq j}^N \frac{1}{r_{ij}^2}, \quad (3.2)$$

where  $r_{ij}$  is the distance between the geometric center of the side-chain atoms of residue  $i$  and the geometric center of the side-chain atoms of residue  $j$ , and  $N$  is the length of the protein. For glycine residues the distance to the  $C_\alpha$  atom was used in lieu of the geometric center of the side-chain.

To calculate Relative Solvent Accessibility (RSA), we first calculated the Accessible Surface Area (ASA) for each site in each protein, via DSSP [46]. We then normalized the ASA values by the theoretical maximum ASA values



found in Table 1 of Tien *et al.* [103]. All WCN and RSA calculations were done on the individual, monomeric protein chain of interest.

All data and analysis scripts required to reproduce the work are publicly available to view and download at [https://github.com/wilkelab/rate\\_variability\\_variation](https://github.com/wilkelab/rate_variability_variation).

### 3.3 Results

We analyzed two distinct data sets. One was a set of 208 diverse enzyme monomers selected from the prior analysis by Yeh *et al.* [111]. The other data set was a smaller set of nine viral proteins from Shahmoradi *et al.* [88]. Note that while the viral data set from Shahmoradi *et al.* [88] includes some viral enzymes, in the following we will use the term “enzymes” to refer specifically to the proteins from the Yeh *et al.* [111] data set.

Homologous sequences for each protein were taken from Yeh *et al.* [111] and Shahmoradi *et al.* [88]. For each protein we made a multiple-sequence alignment using MAFFT [47, 48] on amino-acid sequences. From these alignments we calculated site-specific evolutionary rates using Rate4Site [61]. We measured solvent accessibility for a given residue by its relative solvent accessibility (RSA) (Figure 1.1A). We measured packing density in the protein structures using side-chain WCN (Figure 1.1B). Previous studies have used  $C_\alpha$  WCN when correlating WCN with ER [88, 111, 112]. However, a recent study [59] has shown that calculating WCN using the center of mass of the side-chain results in stronger WCN-ER correlations. Therefore, here we used

side-chain WCN throughout. We also measured the variability in designed sequences. For each protein in the viral data set and for each enzyme less than 200 residues in length we computationally designed 500 sequences using the respective structure as a template. From these sequences we inferred a “design rate” (DR) at each site, calculated as the expected steady-state evolutionary rate for an alignment with the given amino-acid frequencies.

### 3.3.1 Structural Predictors of Evolutionary Rate

To quantify the strength of structure–rate relationships in proteins, we correlated, separately for each protein, structural properties at individual sites with site-specific ER. Unless otherwise noted, we used Spearman correlations throughout. The first structural property that we examined was relative solvent accessibility (RSA). Prior work has shown that RSA has a positive relationship with evolutionary rate [31, 55, 88, 111, 112]. This positive relationship between solvent accessibility and ER was verified in our analysis on the two data sets. Within both data sets, residues that have high RSA evolved faster on average. However, the strength of the relationship between RSA and ER varied between the enzyme and viral protein data sets. The enzymes, on average, had larger RSA–ER correlations with a mean correlation coefficient of 0.55 compared to 0.18 for viral proteins ( $t$  test:  $P = 3.324 \times 10^{-5}$ ) (Figure 3.1A and Table 1).

Next we investigated the relationship between ER and packing density. For both data sets, residues with more contacts evolved slower (Figure 3.1B

and Table 1). This trend was also stronger for enzymes than for viral proteins, with a mean correlation coefficient of -0.63 for enzymes and -0.21 for viral proteins ( $t$  test:  $P = 2.454 \times 10^{-5}$ ).

### 3.3.2 Protein Design as a Structural Predictor

Using protein design to search sequence space, Kuhlman and Baker [50] found that sequences are close to optimal for a given structure (i.e., residues found at a given site are limited for a given structure). This constraint is especially true for buried residues. Given this result, Shahmoradi *et al.* [88] attempted to use site-wise variability in designed proteins as an additional structural predictor of ER [88]. Likewise, here, we used protein design as a third predictor of ER. However, unlike in Shahmoradi *et al.* [88], we did not use design entropy at sites but instead calculated a “design rate” (DR) as our predictor. We calculated this rate by calculating a predicted nonsynonymous substitution rate (dN) from amino-acid frequencies at each site, as derived in Spielman and Wilke [97]. We found that this predicted rate makes similar predictions as does design entropy (not shown). We used design rate here because it is the more principled quantity to compare to ER. For computational feasibility, for the enzyme data set we only designed proteins that were less than or equal to 200 residues in length. This encompassed 32 enzymes. We designed proteins for all the structures in the viral protein data set. Before performing our analysis, we compared the distributions of the strength of structure–rate correlations from the full enzyme data set with that of the

distributions obtained from the 32 proteins. The differences between mean of the distributions were not significant ( $t$  test:  $P = 0.419$  for RSA,  $P = 0.947$  for WCN, Figure 3.2).

In viral proteins, DR had a mean correlation coefficient of approximately -0.02, and in enzymes the mean coefficient of correlation was approximately 0.24 (Figure 3.3 and Table 1). However, for viral proteins this lower mean correlation was slightly misleading because some proteins had positive correlations while others had negative correlations, for a mean near zero (Figure 3.3). In both data sets, design rate was a weaker predictor of evolutionary rates compared to WCN and RSA.

Even though DR did not correlate that strongly with ER, it is possible that it could explain variance in ER not explained by either RSA or WCN. To investigate this possibility, we used DR at sites as a predictor in linear models, either individually or in combination with the two other structural predictors, and calculated the percent variance explained for each model. In general, for both enzymes and viral proteins, design rate was not a good predictor of ER at sites. However, DR, just like RSA and WCN, was better at predicting ER in enzymes than in viral proteins. For a model with design rate as a single predictor, the average  $R^2$  was  $\sim 0.01$  for viral proteins and  $\sim 0.07$  for enzymes (Figures 3.4, 3.5). Including DR as an additional predictor along with RSA and WCN added some additional predictive power for ER in both data sets. For example, the average  $R^2$  of a model with RSA and WCN as predictors for enzymes was approximately 0.37 (Figure 3.4). When we added

DR as an additional predictor, the average  $R^2$  increased to 0.40 (Figure 3.5). This increase in predictive power was observed in the viral data set as well. In summary, although DR was poor predictor of evolutionary rate at sites, it provided a small improvement in model performance, in particular for the enzyme data set.

### 3.3.3 Effect of Divergence of Structure–Rate Relationships

We found WCN, RSA, and DR all to be poor predictors of ER in viral proteins. There could be at least two different explanations for this finding. First, there could be unique structural features found within the viral protein data set that are not in the enzymes as indicated in Tokuriki *et al.* [104]. Second, the viral proteins from Shahmoradi *et al.* [88] may have experienced unique selection pressures (such as immune escape) or different divergence times than the enzymes taken from Yeh *et al.* [111].

We found it unlikely that biophysical differences drove observed differences in the structure–rate correlations between the two data sets. First, any differences between the distributions for mean the WCN of the proteins within the data sets were not significant ( $P = 0.437$  for WCN, Figure 3.6). Differences in the mean RSA of the proteins were significant but the means were extremely similar ( $t$  test:  $P = 0.027$  for RSA, Figure 3.6). Second, the strength of structure–rate correlations was only weakly dependent on the mean WCN or mean RSA of a protein (Figures 3.7, 3.8). Proteins with larger mean RSA had only slightly larger RSA–ER correlations on average and the mean

WCN was not related to the magnitude of structure–rate correlations (Figures 3.7, 3.8).

We next investigated the possibility that differences in the multiple–sequence alignments for the two data sets were driving the differences in predictive power of RSA, WCN, and DR. On average the enzymes have more sequences in their representative alignments. We examined whether this difference was causing the difference in structure–rate correlation strength. We did observe a relationship between the number of sequences and the structure–rate strength. However the strength of this relationship was modest for enzymes ( $\rho = -0.185$ ,  $P = 7.403 \times 10^{-3}$  for WCN–ER and  $\rho = 0.060$ ,  $P = 0.390$  for RSA–ER) and was non-significant for viral proteins ( $\rho = -0.433$ ,  $P = 0.250$  for WCN–ER and  $\rho = 0.633$ ,  $P = 0.076$  for RSA–ER).

The two data sets showed significantly different levels of evolutionary divergence (Figure 3.9). We calculated the divergence for each data set using two quantities: mean root-to-tip distance and mean patristic distance. Root-to-tip distance represents the extent of evolutionary divergence from the data set’s common ancestor to a given sequence. The mean root-to-tip distance for each dataset was calculated as the average branch length, which indicates the number of substitutions, from the root in the tree to each terminal edge (tip) in the tree. Patristic, or pairwise, distance is the sum of branch lengths between two tips in a tree, and indicates how distantly related two sequences are to one another. As with mean root-to-tip-distance, a higher mean patristic distance indicated more evolutionary divergence. The enzyme alignments were much

more diverged than the viral protein alignments ( $t$  test:  $P < 2.20 \times 10^{-16}$  for mean root-to-tip distance and  $P < 2.20 \times 10^{-16}$  for mean patristic distance).

Figure S6 shows structure–rate correlation strengths as a function of divergence (here measured as mean patristic distance). For both RSA–ER and WCN–ER correlations, proteins with MSAs that had higher levels of divergence tended to have higher structure–rate correlations in magnitude. However, the trend between RSA–ER and WCN–ER correlations and mean patristic distance was not very strong ( $\rho = 0.161$ ,  $P = 0.017$  for RSA–ER and  $\rho = -0.117$ ,  $P = 0.086$  for WCN–ER).

Because divergence correlated only weakly with the structure–rate correlations, we hypothesized that overall divergence in an alignment mattered less than did variability in divergence among sites in an alignment. To obtain strong correlations with structural quantities, we need both highly conserved and highly variable sites. To assess the variability in the alignment at each site, we next calculated Shannon entropies at each site. By plotting the variance in entropy among sites against the mean (Figure 3.10A), we found that indeed some alignments had overall high divergence but low variability among sites while other alignments were less diverged on average but had higher variability among sites. Figure 3.10B–F shows specific examples of entropy distributions among sites for individual proteins. For example, consider the protein identified by PDB ID 1G24 (Figure 3.10B). This protein had high mean entropy while maintaining a relatively low variance of entropy. Thus, sites in this protein were uniformly highly variable. Note that the distributions of entropy

varied greatly between proteins even when they were from the same data set (Figure 3.10B–F).

We next plotted structure–rate correlations against the variance in entropy and found strong correlations (Figure 3.11, Spearman’s correlation test:  $\rho = -0.321$ ,  $P = 1.526 \times 10^{-6}$  for WCN–ER,  $\rho = 0.236$ ,  $P = 4.746 \times 10^{-4}$  for RSA–ER). Proteins that had more variance in entropy across sites had larger structure–rate correlations in magnitude. Overall, enzymes were more diverged which in turn resulted, on average, in larger variances in entropy across proteins. The viral proteins were less diverged and as such had lower variances in site variability. However, even for the highly diverged enzymes, correlations with structural quantities were low unless the alignments showed high variation in site variability. Thus, structure–rate correlations are maximized at intermediate levels of divergence, where alignments are sufficiently diverged for a high dynamic range (both highly conserved and highly variable sites are present in the same alignment) but not overly saturated with divergence (so that all sites are highly diverged).

We also investigated the effect of alignment quality on the observed patterns. Highly diverged sequences are more difficult to align, and errors in multiple–sequence alignments may propagate to yield spurious rate inferences at some sites. Such inferences may be partially responsible for the low structure–rate correlations for some proteins. To assess average alignment reliability, we calculated a reliability score using Guidance [73,95] for each multiple sequence alignment. For each alignment, we calculated a column score (CS)



at each site. CS scores range from 0, indicating an unreliably-aligned site, to 1, indicating a highly reliable alignment. We averaged the Guidance CS for each multiple-sequence alignment to obtain a mean Guidance score representing the overall quality of an alignment. All of the viral proteins had scores greater than 0.98, indicating that these alignments had low uncertainty. The enzyme proteins had scores that span a very wide spectrum of quality, from 0 to 1. However, in enzymes, we found that the strength of structure-rate correlations was not correlated with alignment quality (Figure 3.12, Spearman’s correlation test:  $\rho = -0.022$ ,  $P = 0.746$  for WCN-ER,  $\rho = -0.132$ ,  $P = 0.057$  for RSA-ER). This finding suggests that alignment quality is not a significant factor in the observed strength of structure-rate correlations.

As a final test of the effect of divergence on structure-rate correlations, we obtained a series of more diverged viral alignments. Briefly, we used PSI-BLAST to obtain a set of homologous proteins for each of the viral proteins from Shahmoradi *et al.* [88], using the UniProt90 database. This procedure was comparable to the procedure that had been used to assemble the enzyme alignments. Subsequently, we performed the same analysis using these alignments as we did on the other two data sets. Using this new methodology, we only managed to collect sufficient sequences to calculate meaningful evolutionary rates for three of the viral proteins (PDB IDs: 1RD8, 3GOL, and 3LYF). However, even though the data set was small, we could compare it to the other two data sets for consistency. We found that the new viral data set was more diverged than the original viral data set but still less diverged than

the enzyme data set (Figure 3.13). Despite this increased divergence in the new viral data set, the strength of WCN-ER and RSA-ER correlations were similar to the original viral data set. Additionally, the relationship between measures of divergence and the strength of structure-rate correlations was similar for both viral data sets (Figures 3.11, 3.13). Even with the new approach it was difficult to obtain viral alignments with high divergence, which may be responsible for the lower structure-rate correlations still observed.

### 3.4 Discussion

The field of molecular evolution has a long history of attempting to identify the factors that affect the rate at which proteins evolve. At the level of whole-protein rates, some of the factors identified include expression level, interactions with other protein partners [33, 64, 71, 110], and selection for the costs of misfolding [22]. Recently, the emphasis has shifted towards explaining rate variation among sites within proteins, which seems to be driven primarily by biophysical, structural constraints [24, 25, 31, 32, 42, 88, 111, 112].

Among the structural constraints, packing density and relative solvent accessibility have emerged as the two best structural predictors of evolutionary rate [12, 31, 88, 111, 112]. Sites that are on the surface of the protein tend to evolve faster than sites in the protein interior. Similarly, sites that are densely packed and have more contacts tend to evolve slower and exhibit less sequence variability than sites with fewer contacts. However, how strongly these two structural quantities (solvent accessibility and local packing density) correlate

with evolutionary rate at sites remains somewhat unclear.

Here we have examined the relationship between site variability and the strength of structure–rate relationships by performing a direct comparison of the enzyme data set from Yeh *et al.* [111] and the viral proteins from Shahmoradi *et al.* [88]. We have found that both WCN and RSA are significant predictors of ER in enzymes, with 37% of the variation in ER explained (on average) by WCN and 28% explained on average by RSA. In viral proteins, both quantities perform weaker, explaining on average 8% and 7% of variation in ER respectively. Therefore, when analyzed using the same methods the data sets of Yeh *et al.* [111] and Shahmoradi *et al.* [88] both show that WCN performs better than RSA.

In addition to RSA and WCN, we have also considered a third predictor, protein design rate (DR). Protein design had previously been used in Shahmoradi *et al.* [88]. We have found that protein design rate is a much poorer predictor of rates at sites than RSA and WCN are. This result could represent a limitation in current methods of sequence space sampling techniques, limitations in the scoring function used in this study, or it could be that protein design rate does not capture biophysical forces that are predictive of evolutionary rates. For example, Ollikainen and Kortemme [68] published a study that examined the ability of protein design to capture naturally occurring covariation of amino acids at sites. Although flexible-backbone design was able to recapitulate some covariation from natural sequences, not all covariation could be explained by design, indicating that other forces besides

structure could be involved in natural patterns of sequence covariation. Additionally, Jackson *et al.* [43] found that protein design did not recapture some important structure–sequence patterns observed in yeast proteins. Notably, in that study, designed proteins did not exhibit the same relationship between solvent accessibility and site variability observed in natural proteins and hydrophobic residues were often underrepresented in the protein core. These studies underscore the possibility that either current protein design methods are imperfect at mimicking natural structural constraints or that structural constraints do not capture all of the biophysical effects on sequence evolution.

In contrast to the rate predictors in the enzyme data set, for the viral data set, the structural predictors (RSA, WCN, or DR) all performed poorly. We have found that neither differences in structural features (WCN, RSA, or DR) nor differences in evolutionary rates are likely a driving factor in the difference in correlation strength. Therefore, we have investigated the possibility that there are fundamental differences in the two data sets themselves.

We have found that the lack of divergence within the viral proteins of the data set taken from Shahmoradi *et al.* [88] is primarily responsible for the observed low structure–rate correlations. For a protein to have a high structure–rate correlation, there needs to be a high level of variability in divergence among the sites in the multiple–sequence alignment. In other words, a protein must have a combination of sites that are highly conserved and sites that are highly variable. If all sites in a protein are conserved or all sites are saturated with many substitutions, so that there is no variability within

the multiple-sequence alignment, then structure-rate correlations will be low. This combination of highly conserved and highly variable sites will only occur when there is an intermediate level of divergence. This is also why absolute divergence has a much weaker relationship with the strength of structure-rate correlations as compared to variance of entropy. Although it is critical for a data set to have sufficient divergence, it is only a necessary and not a sufficient requirement for strong structure-rate correlations. The enzyme data set of Yeh *et al.* [111] has a variety of proteins with differing levels of divergence and, on average, has MSAs that are more diverged. The intermediate level of divergence in these enzymes results in larger structure-rate correlations.

In addition, variation in selection at sites within a protein can affect the strength of observed structure-rate correlations. Across a protein, structure may differentially affect site variability and hence the strength of structure-rate correlation strength varies. Selection against misfolding can constrain residues within the protein core while selection for key protein-protein interactions [27, 107] and/or against nonspecific protein-protein interactions [54] may impact the variability seen on the protein surface. For example, important binding sites on the surface of the protein might be constrained decreasing the overall variability in variance of site variability. This would result in lower observed structure-rate correlations.

Although proteins as a whole exhibit common selective pressures, depending on the type of protein there might be additional factors that affect rate. Both viral proteins and enzymes exhibit some of the same selective pres-

sures such as selection for stability and pressure to fold and adopt the correct native conformation. Enzymes are used to catalyze chemical reactions and as such have additional constraints such as structural constraints for a proper active site for catalytic function. On the other hand, viruses use their proteins to infect and replicate within their hosts. These proteins are utilized to perform a variety of necessary functions for viral replication such as host cellular entry [2, 80] and nuclear importation [84]. As host immune systems attack these viruses, they evolve to escape from these host mechanisms resulting in signatures of positive selection within these proteins. Because of the differences in selective pressures facing these two protein types there might be different structural constraints on sequence variability and evolutionary rate.

We would like to emphasize that even though the distributions of average WCN and average RSA among proteins are similar for both data sets, there could be other structural differences among the proteins in the two data sets that might affect structure–rate correlations. Our purpose here was not to provide a rigorous, detailed analysis of structural differences among the two data sets. We only examined two obvious structural features (i.e., average packing of residues and average residue solvent accessibility) and showed that they are likely not the cause for the major discrepancy in correlation strengths among the two data sets. More sophisticated structural analyses may identify unique structural features among viral proteins [104], and future research will have to determine whether these features have a measurable impact on structure–rate relationships. Furthermore, our results only apply to the two

data sets discussed. Any additional general conclusions about the impact of divergence on observed structure–rate correlations in other systems would need further study.

### 3.5 Figures

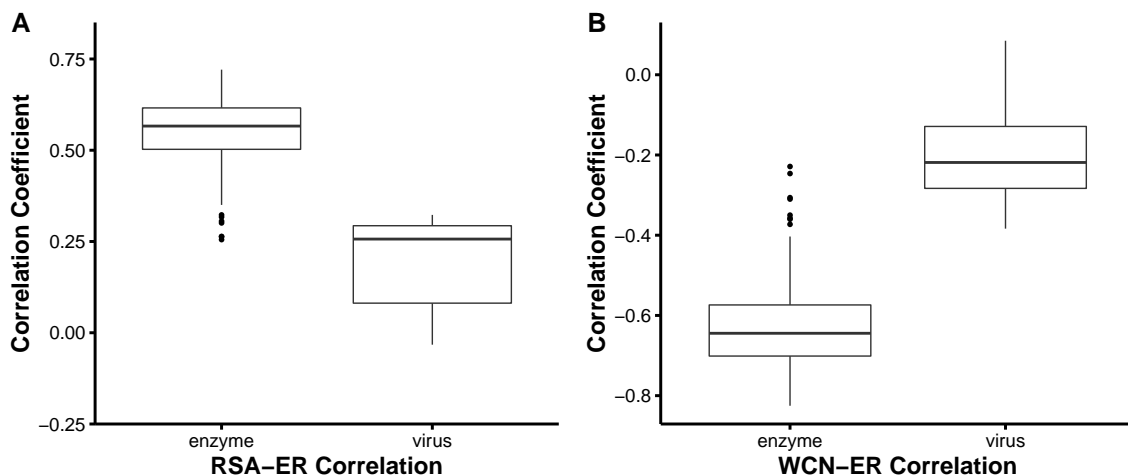


Figure 3.1: Distribution of correlation coefficients between structural properties and evolutionary rate (ER). (A) Spearman correlation coefficients between RSA and ER for the two data sets ( $t$  test:  $P = 3.324 \times 10^{-5}$ ). (B) Spearman correlation coefficients between WCN and ER for the two data sets. For all structural properties, on average, viral proteins show weaker correlations than do enzymes ( $t$  test:  $P = 2.454 \times 10^{-5}$ ).



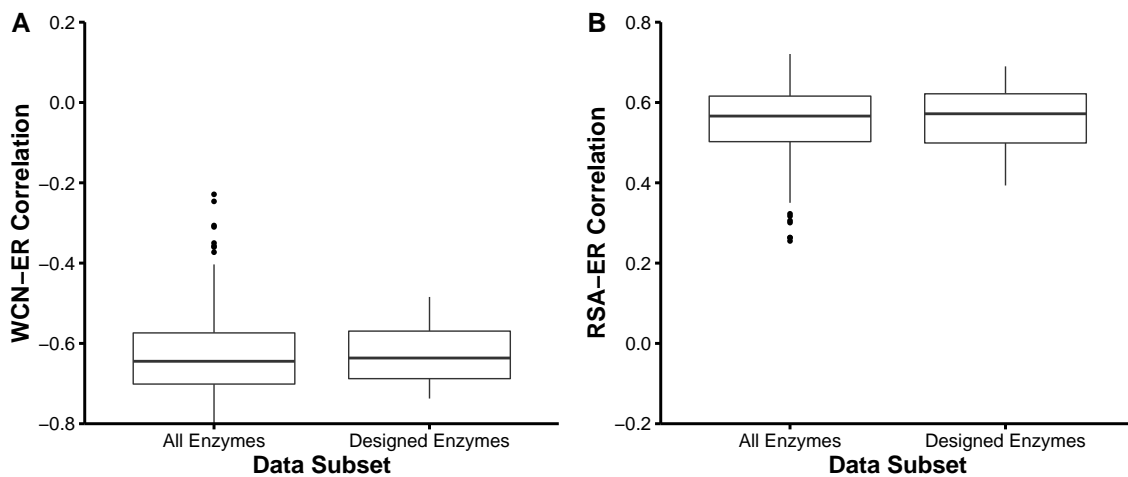


Figure 3.2: Comparison of structure-rate correlations for the full data set of enzymes and the designed set. (A) Comparison of Spearman correlation coefficients for WCN-ER. (B) Comparison of Spearman correlation coefficients for RSA-ER. For both WCN-ER and RSA-ER the mean of the distributions for the designed set of enzymes is the same as that of the full data set of enzymes ( $t$  test:  $P = 0.947$  for WCN-ER,  $P = 0.419$  for RSA-ER).

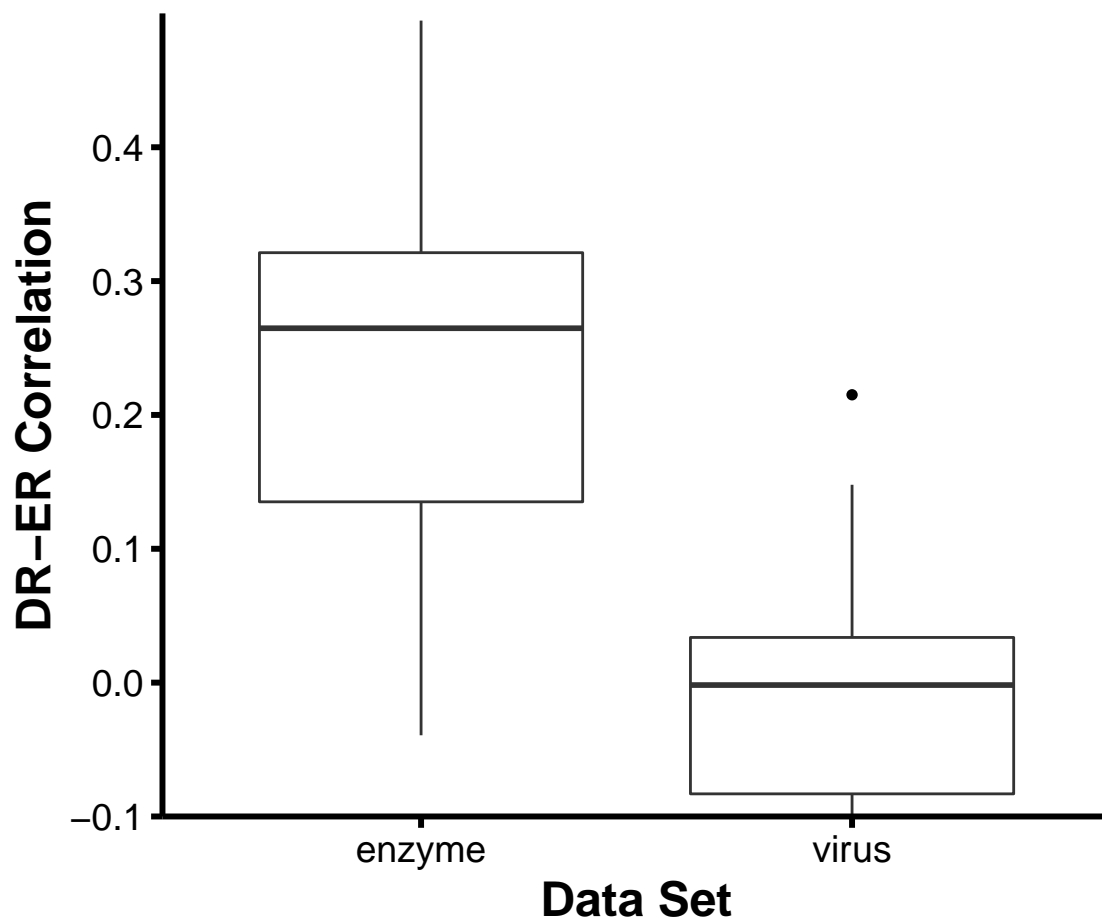


Figure 3.3: Correlation Coefficients of Design Rate and evolutionary rate (ER). Distributions of Spearman correlation coefficients between design rate (DR) and evolutionary rate (ER) for the two data sets. Enzyme proteins have higher correlations on average ( $t$  test:  $P = 7.50 \times 10^{-4}$ ).

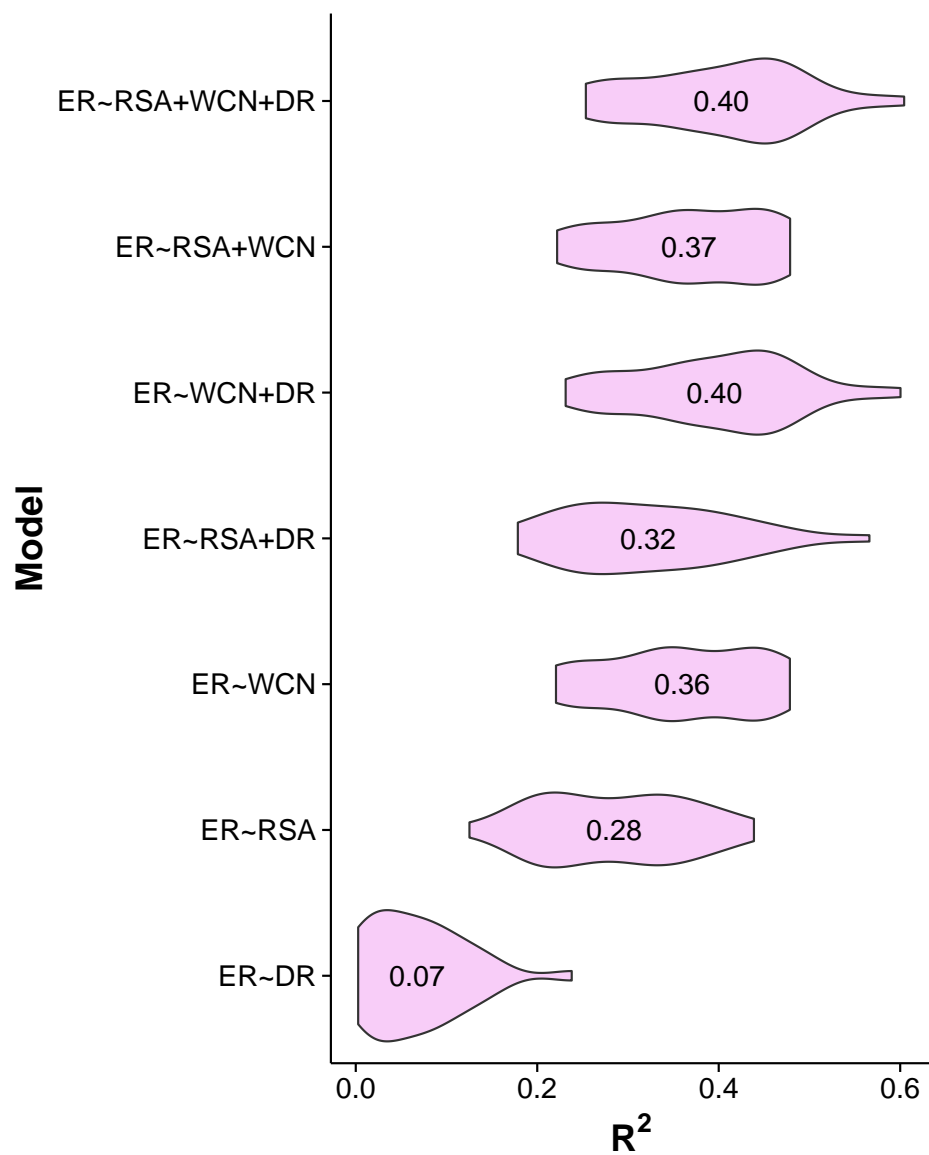


Figure 3.4: Distribution of  $R^2$  for linear models of structural predictors of evolutionary rate (ER) in enzymes. WCN, RSA, DR and all combinations were used as predictors in a linear model with ER at sites as the response. Very little variation in ER can be explained when using design rate (DR) as a single predictor. For enzymes, only 32 proteins were included

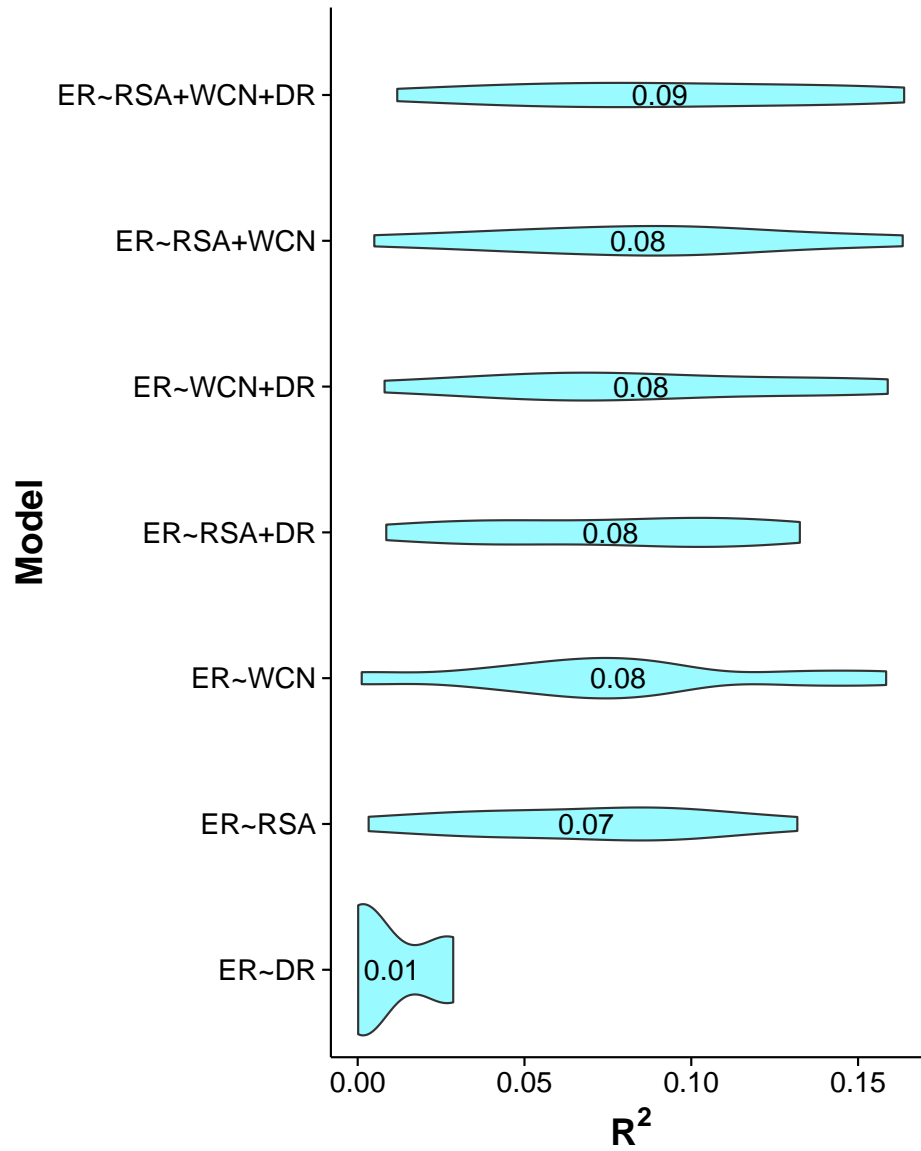


Figure 3.5: Distribution of  $R^2$  for linear models of structural predictors of ER in viruses. WCN, RSA, DR and all combinations were used as predictors in a linear model with evolutionary rate at sites as the response. Very little variation in evolutionary rate can be predicted by RSA, WCN or DR in viral proteins.

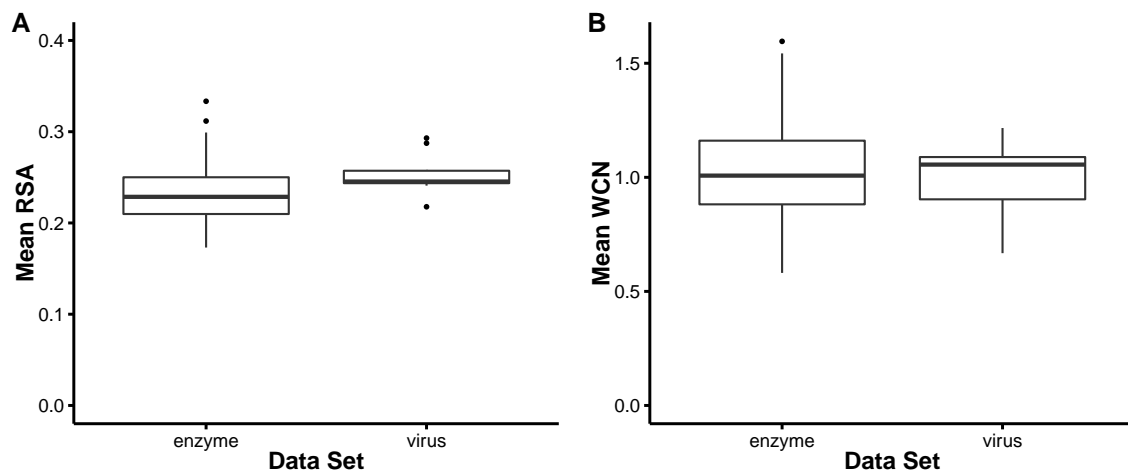


Figure 3.6: Distribution of average structural properties for each protein in the two data sets. (A) Distribution of average RSA. The distribution of average RSA different are very similar for both data sets ( $t$  test:  $P = 0.027$ ). (B) Distribution of average WCN. The distribution of average WCN is the same for both data sets ( $t$  test:  $P = 0.437$ ).

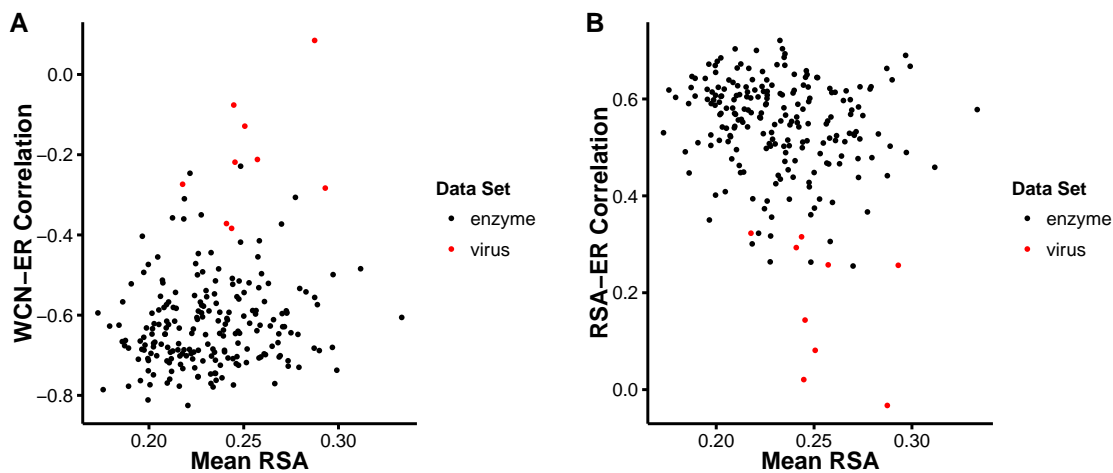


Figure 3.7: Comparison of structure–rate correlations with Mean RSA. (A) Spearman correlations of WCN–ER vs. mean RSA. Proteins with residues that are more exposed on average have slightly larger WCN–ER correlations in magnitude (Spearman’s correlation test:  $\rho = 0.181$ ,  $P = 7.653 \times 10^{-3}$ ). (B) Correlations of RSA–ER vs. mean RSA. Proteins with residues that are more exposed on average also have slightly larger RSA–ER correlations in magnitude (Spearman correlation test:  $\rho = -0.228$ ,  $P = 7.241 \times 10^{-3}$ ).

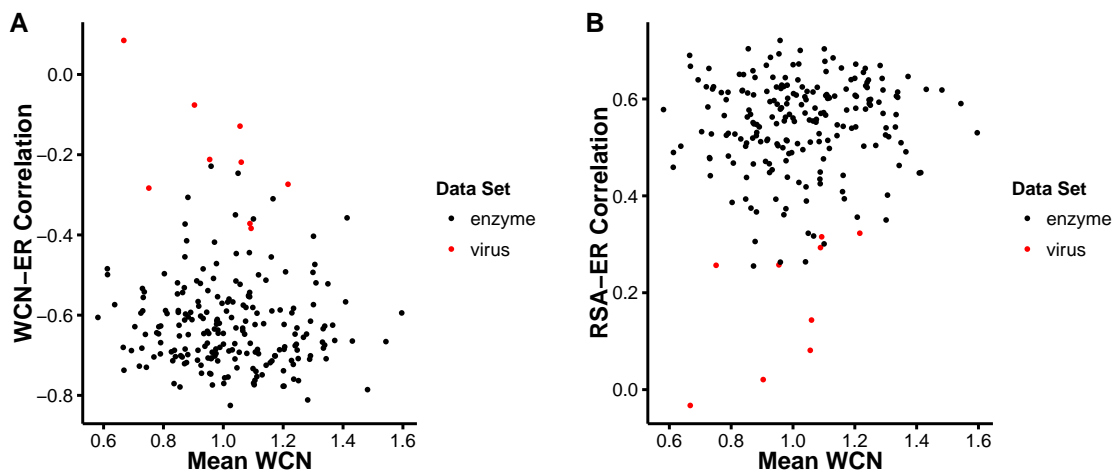


Figure 3.8: Comparison of structure-rate correlations with Mean WCN. (A) Spearman correlations of WCN-ER vs. mean WCN (Spearman correlation test:  $\rho = -0.082$ ,  $P = 0.2283$ ). (B) Correlations of RSA-ER vs. mean WCN (Spearman correlation test:  $\rho = 0.077$ ,  $P = 0.2585$ ). The average WCN of a protein is not related to the strength of structure-rate correlations.

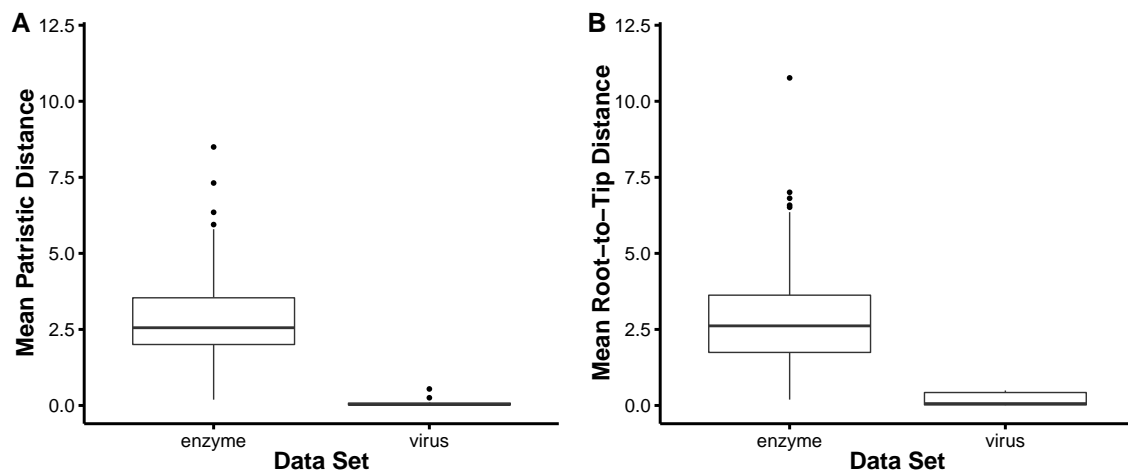


Figure 3.9: Divergence of sequences within the data sets. (A) Distributions of mean patristic distances for sequences in each protein alignment. Enzymes have larger mean patristic distances ( $t$  test:  $P < 2.2 \times 10^{-16}$ ). (B) Distributions of mean root-to-tip distances for sequences in each protein alignment. Enzymes have larger mean root-to-tip distances ( $t$  test:  $P < 2.2 \times 10^{-16}$ ). For both measures of divergence, the proteins within the enzyme dataset are more diverged. Divergence is relatively low between the viral proteins.



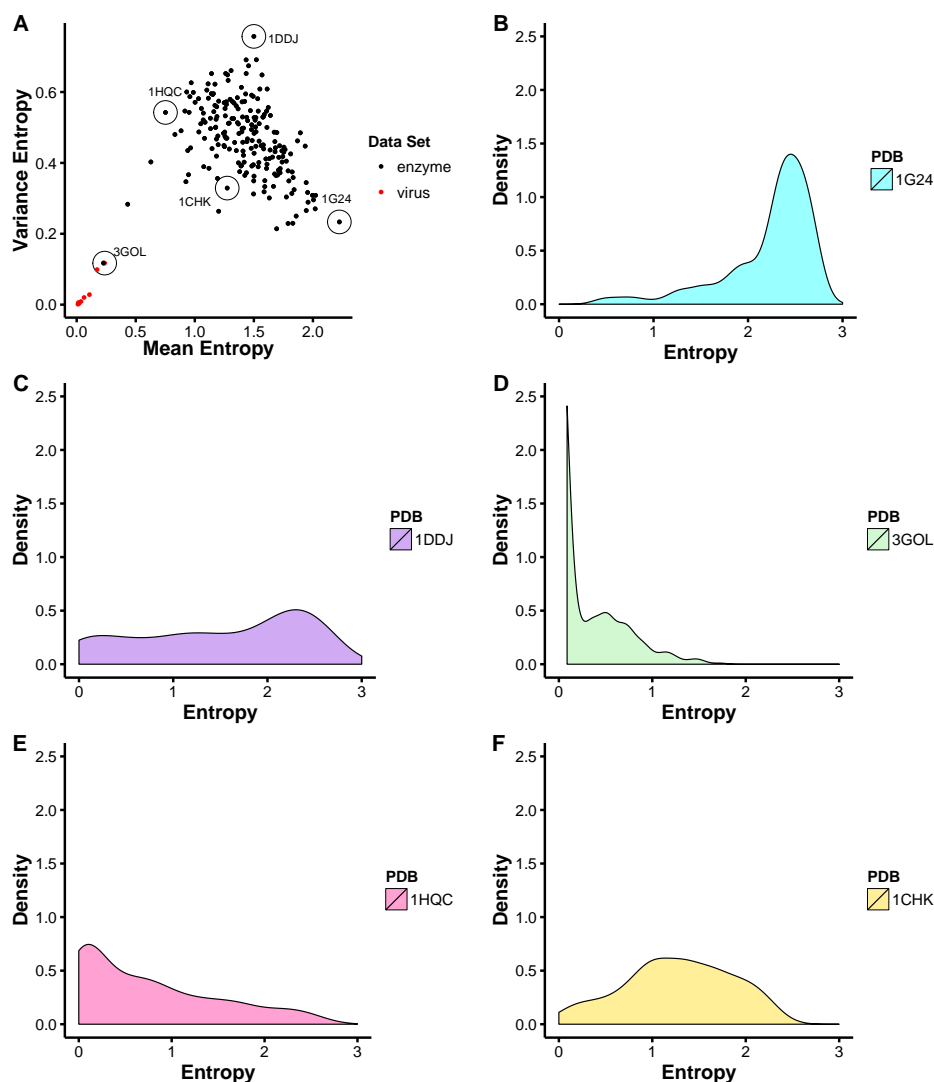


Figure 3.10: Comparison of the mean of entropy and the variance of entropy for individual proteins. (A) Variance in entropy at sites compared against overall mean entropy for each protein. Five different enzymes are highlighted, spanning the range of different combinations of high and low mean entropy and entropy variance. The enzymes are colored in black and the virus proteins are colored red. (B)–(F) Distributions of site-wise entropy values for the five proteins highlighted in A. There are a variety of distributions in site entropy for different proteins. Note: The protein denoted by the PDB ID 3GOL is a viral protein.

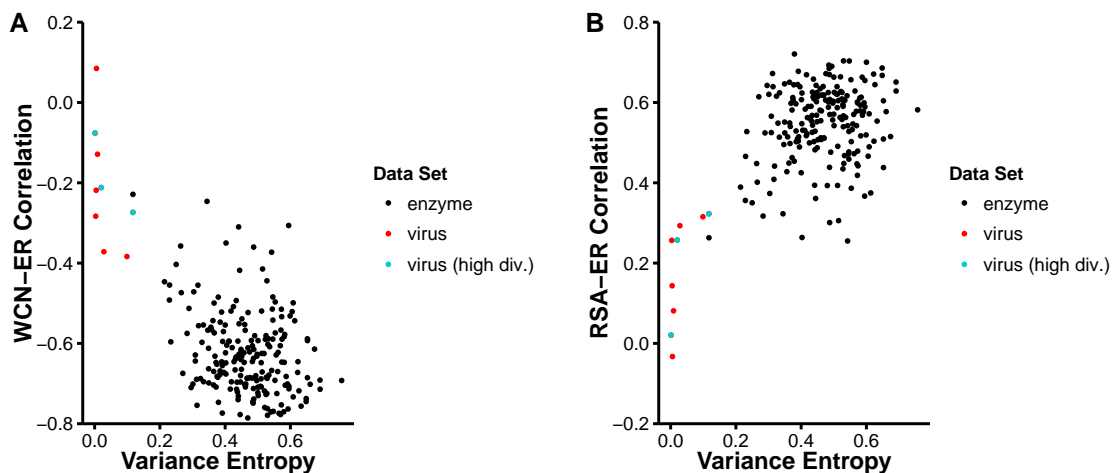


Figure 3.11: Comparison of structure–rate correlations with variance of entropy at sites. (A) Comparison of Spearman Correlation Coefficients of WCN–ER and variance of entropy for proteins. (Spearman’s correlation test:  $\rho = -0.321$ ,  $P = 1.526 \times 10^{-6}$  using only the original protein data sets) (B) Correlations of RSA–ER and variance of entropy for proteins ( $\rho = 0.236$ ,  $P = 4.756 \times 10^{-4}$  using only the original protein data sets). Enzymes are black, the viral proteins with the original alignments are in red, and the viral proteins with the newly collected sequences are in turquoise. Enzymes have more variance in entropy across proteins and have larger structure–rate correlations in magnitude for both RSA and WCN. Virus proteins represented by the newly curated, more diverged alignments (see Methods) have similar structure–rate correlations to the original viral protein data set.

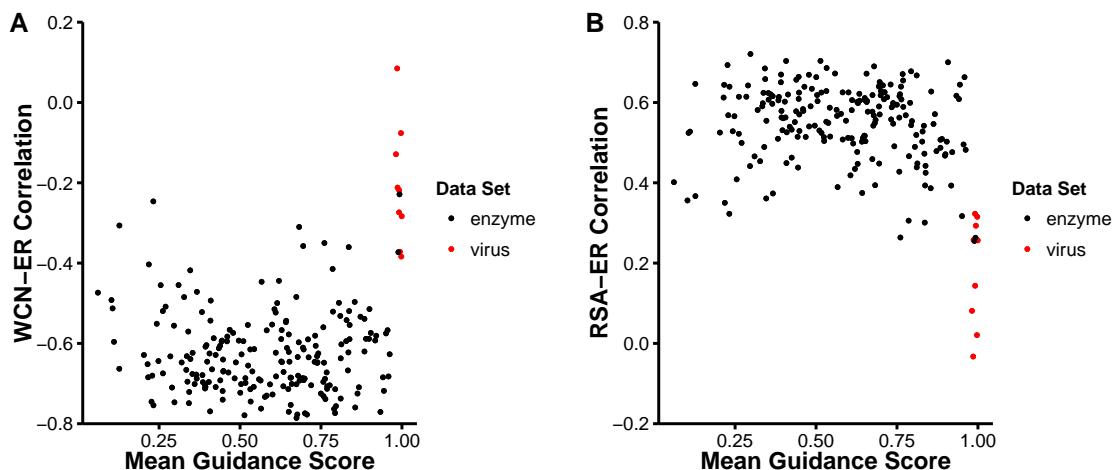


Figure 3.12: Comparison of structure–rate correlations with mean Guidance scores of proteins. (A) Comparison of Spearman correlation coefficients for WCN–ER. (B) Comparison of Spearman correlation coefficients for RSA–ER. Enzymes are black and viral proteins in red. Enzymes have more variation in alignment quality among proteins and have a non-significant relationship between alignment quality and structure–rate correlations (Spearman’s Correlation test:  $\rho = -0.023$ ,  $P = 0.746$  for WCN–ER and  $\rho = -0.132$ ,  $P = 0.057$  for RSA–ER). For viral proteins there is no significant relationship between alignment quality and structure–rate correlations ( $\rho = -0.633$ ,  $P = 0.076$  for WCN–ER and  $\rho = 0.317$ ,  $P = 0.410$  for RSA–ER).

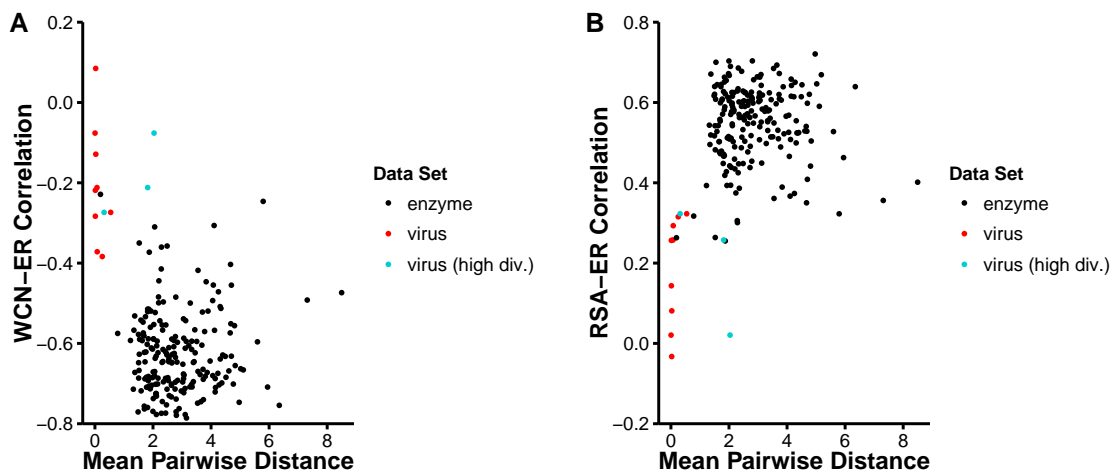


Figure 3.13: Comparison of structure–rate correlations with divergence. (A) Spearman’s correlation test:  $\rho = -0.117$ ,  $P = 0.086$  for WCN–ER. (B) Correlations of RSA and ER vs. mean pairwise distance. Enzymes are black, the viral proteins with the original alignments are in red, and the viral proteins with the newly collected sequences are in turquoise. Proteins that are more diverged (as represented by mean pairwise distance) have stronger RSA–ER correlations (Spearman’s correlation test:  $\rho = 0.161$ ,  $P = 0.017$ ).

### 3.6 Tables

Table 3.1: Averages of Spearman correlation coefficients between structural properties and evolutionary rate (ER). The structural properties analyzed are RSA, WCN, and predicted rate of designed proteins (DR). The analysis was performed on two data sets, one comprised of 208 enzyme monomers and comprised of nine viral proteins. Structure–ER correlations are higher in absolute magnitude in enzymes.

Dataset	$\langle \rho_{\text{ER-WCN}} \rangle$	$\langle \rho_{\text{ER-RSA}} \rangle$	$\langle \rho_{\text{ER-DR}} \rangle^a$	$\langle \rho_{\text{ER-WCN}} \rangle^a$	$\langle \rho_{\text{ER-RSA}} \rangle^a$
Enzyme	−0.626	0.549	0.240	−0.625	0.561
Virus	−0.207	0.184	−0.022	−0.207	0.184

Note:

<sup>a</sup>Correlations coefficients calculated using the 32 enzyme proteins and nine viral proteins for which there were designed sequences.

## Chapter 4

# Computational prediction of the tolerance to amino-acid deletion in green-fluorescent protein

This work is in review in the journal *PLOS ONE*.<sup>1</sup>

### 4.1 Introduction

Proteins must fold into the correct shape in order to function properly. Mutations in DNA can cause amino-acid substitutions that can have a critical effect on the ability of a protein to fold and, hence function, properly. The result is that the evolution of protein sequences is constrained by protein structure. In fact, there have been several works that have examined the effect of protein structure on the evolution of proteins [56, 90, 109, 115]. However, most of these recent studies have only considered the effects of structure on amino-acid substitutions. How protein structure constrains the functional consequences of deletions within proteins is less understood.

Recently, Arpino *et al.* [5] mapped the functional consequences of dele-

---

<sup>1</sup>E. L. Jackson, S. J. Spielman, and C. O. Wilke. Computational prediction of the tolerance to amino-acid deletion in green-fluorescent protein. In review. C. O. Wilke helped to design the project and write the manuscript. S. J. Spielman helped write the manuscript.

tions in enhanced green fluorescent protein (EGFP). The authors found that, on average, functional mutants were largely found in unstructured loop regions as opposed to the highly structured beta sheets and alpha helices. In addition, non-functional mutants were more likely to have deletions in residues that had lower solvent accessibility. The results from this study indicate that structure may play a critical role in whether a deletion at a given residue will be tolerated. However, this hypothesis was not rigorously tested. Here we carefully examine whether structural constraints may impact the relationship between deletions and function in EGFP. We do so by determining whether structural properties can be used to determine whether or not a deletion will result in a functional protein.

In previous studies on the structural predictors of evolutionary rate, both solvent accessibility and local packing density (LPD) have emerged as two prominent structural predictors of evolutionary rate at sites. Solvent accessibility is a measure of the amount of solvent that a given residues comes into contact with. Solvent exposure is often measured by relative solvent accessibility (RSA) (Figure 1.1A). Sites with an RSA of one are completely exposed to solvent. These sites are found on the surface of the protein. Sites with a RSA of zero are buried and do not interact with solvent at all. In proteins, RSA has a positive relationship with evolutionary rate. Residues that have high RSA evolve more quickly [12, 31, 32, 38, 65, 69, 85].

Local packing density (LPD) measures how tightly packed a given residue is within the three-dimensional structure of the protein. Residues

that are tightly packed have many given neighbors within a structure. LPD is often measured by weighted contact number (WCN) (Figure 1.1B). The WCN of a residue is calculated by the sum of the inverse of the radius between that residue and all of its neighbors. In proteins, WCN has a negative relationship with evolutionary rate. Residues that have high WCN (i.e., are tightly packed within the three-dimensional structure) evolve more slowly [42, 55, 111, 112]. Since both of these properties have shown to be significant predictors of evolution at sites, we use both of these properties along with the secondary structure (SS) of a residue to predict whether a deletion will yield a functional protein.

In addition, we use protein design to explicitly model and computationally score each mutant and use these scores as a predictor of functional status. We find that all three simple structural properties (i.e., WCN, RSA, SS) are significant predictors of whether a deletion will be tolerated at a given site. We also find that using protein design to explicitly model deletions within the structural context results in better predictive power than using either RSA or SS as a single predictor of functional status. WCN is the best single predictor of whether a deletion will result in a functional protein. However, using computational scores from designed proteins in a model with other structural predictors does result in improved predictions. Overall, protein structure appears to be a crucial factor in determining tolerance to deletion in proteins. This implies that lack of function due to deletions is, at least in part, a result of structural disruption leading to incorrect folding.



## 4.2 Materials and Methods

### 4.2.1 Functional Data for Mutants

For our analysis we used the crystal structure of EGFP (PDB ID: 4EUL). All functional data corresponding to each mutant was taken from Arpino *et al.* [5]. Briefly, using a transposon-mediated directed evolution trinucleotide deletion experimental approach [44,91], Arpino *et al.* [5] made trinucleotide deletions within the DNA sequence of *EGFP*. Due to their approach, mutations could span multiple codons and mutants could result in proteins with double deletions, single deletions, or a deletion and a non-synonymous substitution. After making the deletions, Arpino *et al.* selected a set of mutants and separated and described each mutant as functional or non-functional. Functional mutants were those that still resulted in *E. coli* that exhibited the fluorescent green phenotype when screened with UV light. Hereinafter we call deletions tolerated if the protein still allows the *E. coli* to exhibit the fluorescent phenotype when under UV light and non-tolerant deletions are those that do not. After selecting, assaying, and sorting each mutant, the final data set was comprised of 87 unique mutants, 42 of which were functional and 45 of which were non-functional.

We only chose a subset of the original mutants. If a mutant had two unique deletions at the nucleotide level but resulted in the same translated product we only kept one mutant representing the final translated protein sequence. This resulted in the removal of one mutant. We removed four non-functional mutants that resulted in the introduction of stop codons into the

sequence. We also removed mutants that had mutations in the N-terminus, C-terminus, or the chromophore. Lastly, we removed mutants with two deletions. In total our data set was comprised of 72 mutants, 34 of which were functional and 38 of which were non-functional.

#### 4.2.2 Calculation of Structural Properties

We calculated the solvent accessibility (ASA) for each deleted residue using DSSP [46]. We normalized ASA to obtain relative solvent accessibility (RSA) by normalizing each residue by the maximum solvent accessibility for each residue type (from Table 1 in Tien *et al.* [103]). We calculated the side-chain weighted contact number (WCN) as defined by Marcos and Echave [59]. Side-chain WCN is defined as:

$$\text{WCN}_i = \sum_{i \neq j}^N \frac{1}{r_{ij}^2} \quad (4.1)$$

where  $r_{ij}$  is distance between the geometric center of the side-chain atoms of residue  $i$  and the geometric center of the side-chain atoms of residue  $j$  in a protein that is  $N$  residues long. For glycine residues, the  $C_\alpha$  atom was used instead of the geometric center. Although previous studies have often used WCN calculated with  $C_\alpha$  atoms, recent work has shown that using the center of mass of the entire side-chain results in stronger correlations between WCN and evolutionary rate [59]. Therefore we use side-chain WCN throughout our study. We predicted the secondary structure (SS) for each deleted residue using Psipred [11, 45].

### 4.2.3 Structural Modeling

For structural modeling we used the crystal structure of enhanced green fluorescent protein (EGFP) as the wildtype from Arpino *et al.* [6] (PDB ID: 4EUL). EGFP is an engineered mutant of the naturally occurring protein, green fluorescent protein (GFP), found in *Aequorea victoria*, a jellyfish. The structure of EGFP is formed by a beta-barrel composed of eleven beta sheets (Figure 4.1). The chromophore that results in the green fluorescent phenotype is in the middle of an alpha-helix that is housed in the middle of this beta-barrel. Before modeling we had to design an EGFP structure without the chromophore molecule. We used RosettaModel [41] to do this. First we removed all non-protein and non-chromophore atoms from the crystal structure. Second, we deleted the chromophore molecule from the structure. We renumbered the structure such that the first residue was numbered one and all following residues were numbered sequentially. Before modeling, we then used the relax protocol [19,67] in Rosetta to optimize and re-pack the side-chains. We created 100 structures in Rosetta using the relax protocol and selected the best model as the template for design based on total score, with the best score being the most negative. We used the following commands for the relax protocol.

```
-database /path/to/rosetta_database  
-s 4EUL_no_cro.pdb  
-ignore_unrecognized_res
```

```
-use_input_sc  
-constrain_relax_to_start_coords  
-nstruct 100  
-relax:fast  
-overwrite  
-out:file:scorefile relax_scorefile.fasc  
-out:path:pdb ./output_pdbs/
```

We used Psipred [11, 45] to predict the secondary structure of EGFP. Using this secondary structure information we used RosettaRemodel to insert the three chromophore-forming residues (i.e., Thr65-Tyr66-Gly67) into the structure where the original chromophore was. Based on the secondary structure information, we built the insert with a helical backbone. In addition to inserting these three residues, we also designed two residues on either side of the insertion to accommodate any major structural changes created by the insertion. The following flags were used for this remodeling procedure.

```
-database /path/to/rosetta_database  
-s 4EUL_no_cro_relaxed.pdb  
-remodel:blueprint design_blueprint.txt  
-run:chain A  
-num_trajectory 5  
-save_top 5  
-ex1
```

```
-ex2  
-extrachi_cutoff 1  
-use_input_sc  
-linmem_ig 10  
-remodel:use_pose_relax  
-out:file:scorefile design_protocol.fasc  
-out:path:pdb ./output_pdbs/  
-remodel:hb_srbb 1.0  
-overwrite
```

We made five structures using RosettaRomodel. We chose the best candidate from these structures based on overall score and manual inspection. This structure served as our wild-type template for modeling the mutants. We used Modeller [28] to model each of the 72 selected mutants from Arpino *et al.* [5]. First, for each target mutant we created a target-template sequence alignment by aligning the sequence for the mutant to the wild-type template structure sequence for EGFP using the software MAFFT [47,48]. We specified the “auto” flag in MAFFT to specify the optimal alignment algorithm for each alignment. This designed and relaxed EGFP with no chromophore served as our template structure. We used Modeller to model 25 homology modeling structures for each mutant.

We relaxed the resulting 25 modeled structures for each mutant using the relax protocol in Rosetta. Below are the flags used for this relaxation procedure.

```
-database /path/to/rosetta_database
-l pdb_list.txt
-ignore_unrecognized_res
-use_input_sc
-constrain_relax_to_start_coords
-flip_HNQ
-no_optH false
-nstruct 4
-relax:fast
-overwrite
-out:file:scorefile scorefile.fasc
-out:path:pdb ./output_pdbs/
```

For each protein we generated four relaxed structures. As a result there were 100 final structures that corresponded to each mutant (25 Modeller models x 4 relaxed structures each = 100 structures). We used the mean Rosetta score for each of the 100 structures as a predictor of deletion tolerance. A schematic of the entire RosettaRemodel protocol is visualized in Figure 4.2. All scripts and data can be found at [https://github.com/wilkelab/EGFP\\_deletion\\_prediction](https://github.com/wilkelab/EGFP_deletion_prediction).

#### 4.2.4 Statistical Analysis of Functional Status

We used two different machine learning approaches to predict functional status using structural predictors, logistic regression and a support vector

machine. For each approach we used the same models. We used WCN, RSA, SS, and “mean score” as our structural predictors. Mean score is the average Rosetta score from each of the 100 modeled structures for each mutant. For each model, we used either the structural predictor by itself or in combination with others. We tried all combinations of the structural predictors. For the logistic regression analysis we used the “glm()” function in R [79] with “family = binomial” to specify the logic link function.

We implemented a supervised support vector machine algorithm using the “e1071” [63] package in R [79]. For our support vector machine we used a radial basis kernel with default parameters (i.e.,  $\gamma = (1/\text{data dimension})$  and the default cost of constraint violation (C) of one).

For each approach we used 10-fold cross validation for each model. Briefly, for each data set, all points in the nine other data sets were used as a training data set to train a model and then that trained model was used to make predictions for the remaining mutants. ROC curves for each model were obtained by pooling all of the predictions from the 10 test data sets and plotting the true positive rate versus the false positive rate for each model. The true positive rate represents the number of mutants that were correctly identified as tolerated. This was calculated as the number of mutants identified as tolerated by the model divided by the number of known tolerated mutants. The false positive rate represents the percent of mutants that are incorrectly identified as tolerated. This was calculated by dividing the number mutants that were falsely identified as tolerated by the model divided by the number

of known non-tolerated mutants. We used the Area Under the Curve (AUC) value of each of the ROC curves from the predicted data points to assess the predictive ability of each model. We call this the “cross-validated AUC”. We repeated the cross validation procedure 100 times and calculated a mean cross-validated AUC for each model. A mean cross-validated AUC value of 0.5 implies random prediction. Any value over 0.5 implies better than random prediction by the model and an AUC of 1 implies perfect prediction.

### 4.3 Results

Here we attempted to examine the effect of protein structure on deletions using EGFP as a test case. To examine the relationship between protein structure and tolerance to deletion we took functional data for 72 EGFP mutants from Arpino *et al.* (2014) [5]. A given structure was tolerated if after the deletion the protein was still functional. For each mutant protein we measured RSA, WCN and SS for the deleted residue. We predicted the secondary structure (SS) for each mutant using Psipred [11]. In addition, for each of the 72 mutants, using RosettaRemodel and Modeller, two computational modeling techniques, we used the wild-type structure as a template to design structures for each deletion (or deletion followed by a substitution). For each mutant we designed 100 structures and calculated the mean Rosetta score. We called the mean of the Rosetta scores for these 100 designed structures the “mean score” and used it as an additional structural property.



### 4.3.1 Variation in structural properties between non-tolerated and tolerated deletions

In this data set we have found that residues that had deletions that resulted in non-functional proteins had lower RSA (Figure 4.3A,  $t$  test:  $P = 1.030 \times 10^{-3}$ ). Likewise, residues that resulted in non-tolerated deletions had higher WCN on average (Figure 4.3B,  $t$  test:  $P = 2.998 \times 10^{-7}$ ). In Arpino *et al.* [5], the authors noted that most of the functional deletions were present in unstructured loop regions and that most of the non-tolerated deletions were found in the highly structure beta sheets. We verified this trend (Figure 4.3C). In addition to RSA, WCN, and SS, we used protein design and homology modeling to explicitly model each mutant. For each mutant we created 100 models (for protocol see methods) and calculated a "mean score", which is the average of the computational scores from all 100 models. We found that, on average, mutants that were classified as tolerated had lower (more negative) scores (Figure 4.3D,  $t$  test:  $P = 2.084 \times 10^{-6}$ ). This suggests that computational scores of designed proteins are somewhat indicative of deletion tolerance. Overall, we found that structural properties vary between tolerated and non-tolerated deletions.

### 4.3.2 Functional Classification Prediction

The systematic variation in structural properties between tolerated and non-tolerated deletions suggested that structure was indeed a viable metric for predicting tolerance to deletions. Therefore we attempted to use these struc-

tural properties to directly predict the functional status of a given mutant. First we attempted to use solvent accessibility (as measured by RSA) and local packing density (as measured by WCN) to predict tolerance to deletion. We used RSA and WCN as single predictors of functional status in two logistic regression models. We used 10-fold cross validation for each model (see Materials and Methods for details). WCN is a much better predictor of functional status than RSA (a mean cross-validated AUC of 0.820 for WCN versus 0.681 for RSA on the test mutants). Nevertheless, both WCN and RSA were significant predictors of function (Table 4.1) and both models made predictions that are significantly better than random chance (i.e., mean cross-validated AUC greater than 0.5).

We next attempted to see if the secondary structure (SS) of a residue could be used to predict the effect of deletion on functional status. Secondary structure for a given residue had three possible values: beta sheet (sheet), alpha helix (helix), or loop (loop). The mean cross-validated AUC for a model with SS as a single predictor was 0.706. This implies that the location of residue in a given structured or unstructured region is a better predictor of tolerance to deletion than the solvent accessibility of a residue, but a weaker predictor than WCN. Interestingly, the mean AUC when using the entire data set to train the model (the model AUC) with SS as a single predictor was much higher than that of the corresponding mean cross-validated AUC value. For WCN and RSA the AUC value for the model was much more similar to the cross-validated AUC value. This suggests that models SS are more sensitive to

the training set used and that WCN and RSA are more consistent predictors when developing a model predicting tolerance to deletions.

As a final predictor we used the average score of the 100 designed structure scores (mean score) as a single predictor of functional status. Although mean score was a better predictor than RSA and SS, WCN was still the best predictor of tolerance to deletion of a residue (Table 4.1). However, it is possible that modeling each individual mutant might be useful for developing a more predictive model of functional tolerance to deletion by adding it as an additional predictor to a model with RSA, WCN, or SS. Therefore we built additional models that used RSA, WCN, SS, and mean score in various combinations. Indeed, using mean score in combination with other structural predictors did increase predictive ability (Table 4.1). The overall best model for predicting functional status was the model with RSA, WCN, and mean score as predictors with a mean cross-validated AUC value of 0.902. This model was significantly better than the next best model, the model with RSA, WCN, and SS as predictors, which had a mean cross-validated AUC value of 0.885 ( $t$  test:  $P < 2.2 \times 10^{-16}$ ). In fact, four of the top six logistic regression models scored by mean cross-validated AUC had mean score as a predictor.

We used a second machine learning approach to predict functional status using the same four predictors. We used a support vector machine (SVM) to predict functional status using WCN, RSA, SS, and mean score as predictors. Once again we used 10-fold validation of our models and used mean cross-validated AUC again as a measure of prediction accuracy. Except in the

case of the model with WCN, SS and mean score as predictors of functional status, the mean cross-validated AUC value for all models were higher when using logistic regression (Figure 4.5). However, differences between the two approaches were minor and most of the results from the SVM analysis largely agreed with those from the logistic regression analysis (Table 4.2). Although in a slightly different order (the second and third scoring models are reversed), the top six scoring models are the same as those in the logistic regression analysis. The model with RSA, WCN and mean score as predictor was once again the best scoring model. With a mean cross-validated AUC value of 0.873, it was significantly better than the second best model, the model with all four predictors, that had a mean cross-validated AUC value of 0.871 ( $t$  test:  $P = 4.163 \times 10^{-6}$ ).

It appears that explicitly modeling of mutants and adding the derived mean score as a predictor to a model with other structural properties did indeed increase predictive ability of functional status. However, it is possible that adding mean score as an additional predictor increased predictive ability by predicting noise. Therefore we performed a principal component analysis of the structural predictor variables and regressed the response (i.e., functional status) on the components. Most of the variance in the data could be explained by PC1 and the non-functional and functional mutants largely separated along PC1 (Figure 4.4A). By plotting the loadings of each of the structural variables on the principal component axes we could see which variables were related in terms of the amount of variation that they explained. Most of the struc-

tural variables generally loaded on PC1 (except for beta sheet) but differed in whether they loaded negatively or positively on PC2 (Figure 4.4B). Mean score loaded positively on PC1 and negatively on PC2. RSA, WCN and all secondary structure elements slightly differed in terms of how they loaded on the two axes. This implies that mean score increased predictively ability because it explained additional variation that was not captured by the other structural variables. Therefore it appears that explicitly designing proteins for each mutant does help provide more accurate predictions of whether a given deletion will be accepted.

## 4.4 Discussion

Here we performed a systematic study to investigate how protein structure affects tolerance to deletion by using structural properties to predict tolerance to deletions using enhanced green fluorescent protein (EGFP) as a model protein. We first determined the extent to which WCN and RSA could be used to predict whether a given deletion would be tolerated (i.e., result in a functional protein product) or non-tolerated. Both RSA and WCN could be used to predict functional status with WCN being the best predictor. In this data set, we have found that residues that had deletions that resulted in non-functional proteins had lower RSA. In addition, deletions that resulted in non-functional proteins often had higher WCN. These two trends are consistent with current evidence that residues on the surface of proteins evolve more quickly and that residues that are densely packed evolve slower [31, 42, 55, 88, 111, 112]. This

suggests that deletions at sites in proteins might undergo similar selective constraints as do amino-acid substitutions at sites.

In a previous study, Arpino *et al.* (2014) found that structured areas (i.e., beta sheets and alpha helices) were enriched for residues that were not tolerant to deletions. Deletions that occurred in disordered regions (ex. loops) were often more tolerated. This is in line with earlier work that found that most indels are found in turns, coils, or disordered loops and are much rarer in the more structured alpha helix and beta sheet regions of proteins [72,105]. Therefore we added SS as an additional predictor. We found that WCN and RSA were both better predictors of functional status after deletion. However, in this work we only focused on mutations that resulted in single residue deletions. Although the majority of deletions found in proteins are between 1-5 residues in length, there have been some deletions that are much longer in length [15, 72, 102]. Therefore in order to truly examine the relationship between secondary structure and the functional status of deletions we would need to include deletions that are longer in length in future work.

In addition to using these simple structural properties to predict tolerance to deletion we also used protein design to explicitly model each mutant and used average model score as a predictor of functional status after deletion. We have found that while the average score of designed models was a good predictor of functional status, WCN was still a better predictor of functional status. However, we used mean score in combination with the other three structural predictors in a series of models to determine whether or not

explicitly modeling each mutation might result in improved models when used in combination with the other predictors. While WCN was found to be the best single structural predictor of deletion, computational modeling generally resulted in better prediction capability no matter what method was used for prediction (i.e., logistic regression or SVM). This suggests that the explicit modeling is adding some information that allows for predicting tolerance to deletions when using machine learning techniques.

In general RSA and WCN, both simple measures of protein structure that have been previously, are both good predictors of deletion tolerance in structures. Since both of these measures are also implicated in the constraints on substitutions at sites, it is likely that these two measures can be generalized to all proteins in terms of their role of constraining deletions at sites. However, secondary structure may not be a good structural predictor of deletion tolerance in all proteins. Although previous research has implicated that deletions at residues in beta sheets are often non-tolerated, our PCA analysis suggests that most of this effect may be attributed to packing density and solvent accessibility. In order to untangle the effect of secondary structure future work using more proteins will need to be performed. Overall, our work suggests that structure does constrain tolerance to deletion at sites. However, here we only studied one protein, EGFP. More studies that incorporate many proteins will be necessary to fully understand the role structure plays in the functional effects of deletions in proteins.

## 4.5 Figures

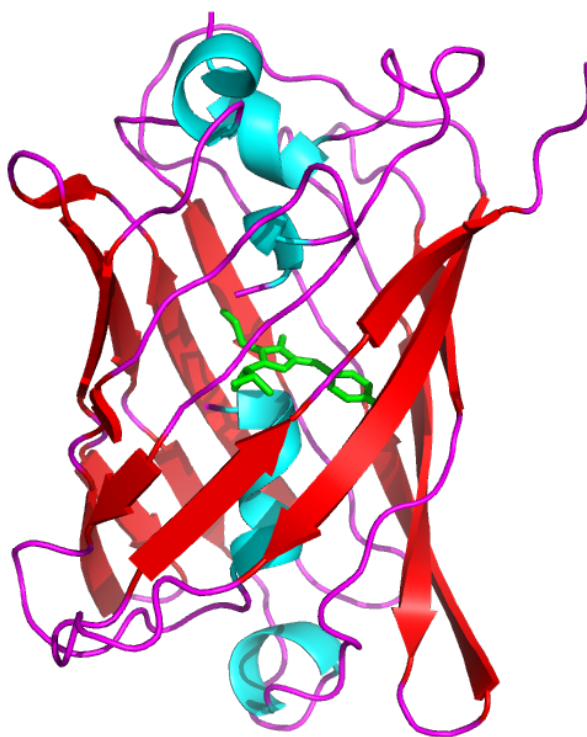


Figure 4.1: Structural Representation of enhanced GFP (EGFP). Secondary structure elements are colored. Beta sheets are colored in red, alpha helices are colored in cyan and loops are in magenta. The chromophore responsible for florescence is colored in green. The structure of EGFP is formed by a beta-barrel composed of eleven beta sheets. The chromophore that results in the green fluorescent phenotype is in the middle of an alpha helix that is housed in the middle of this beta-barrel.



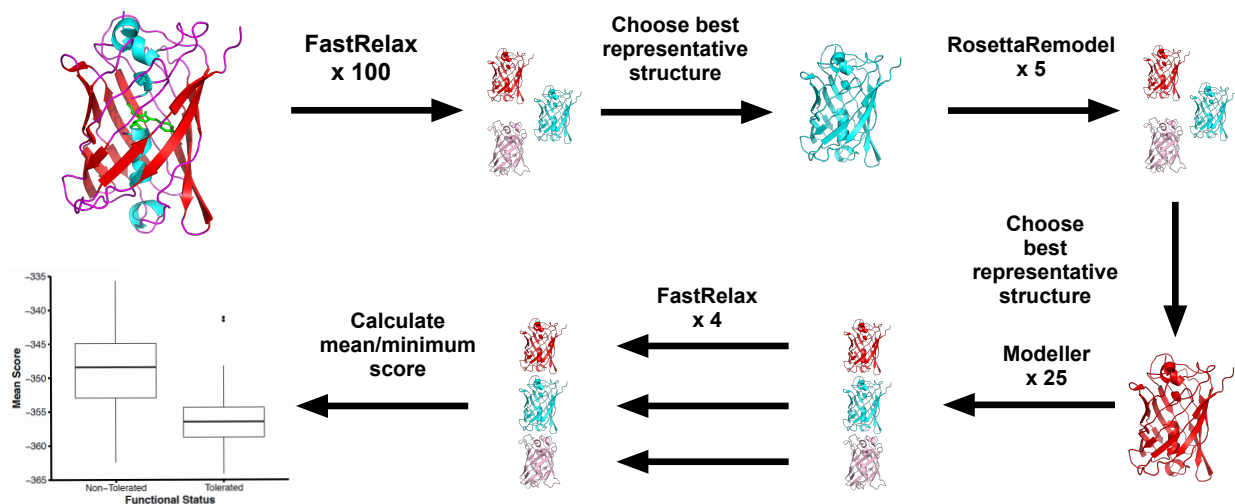


Figure 4.2: Visualization of the computational modeling pipeline. The colors represent variation between structural models produced by each protocol. After removing the chromophore, we used the relax protocol in Rosetta to optimize the structure. We chose the lowest scoring model from the 100 created models as our modeling template. We used RosettaRemodel to model the EGFP structure without the chromophore. Using the lowest scoring model from the protocol as our template, we used Modeller to model 25 mutants for each deletion mutant. We used the relax protocol to create four optimized structures for each of the 25 homology models for each mutant. We took the mean of the 100 models and used this as a predictor of functional status for each mutant.

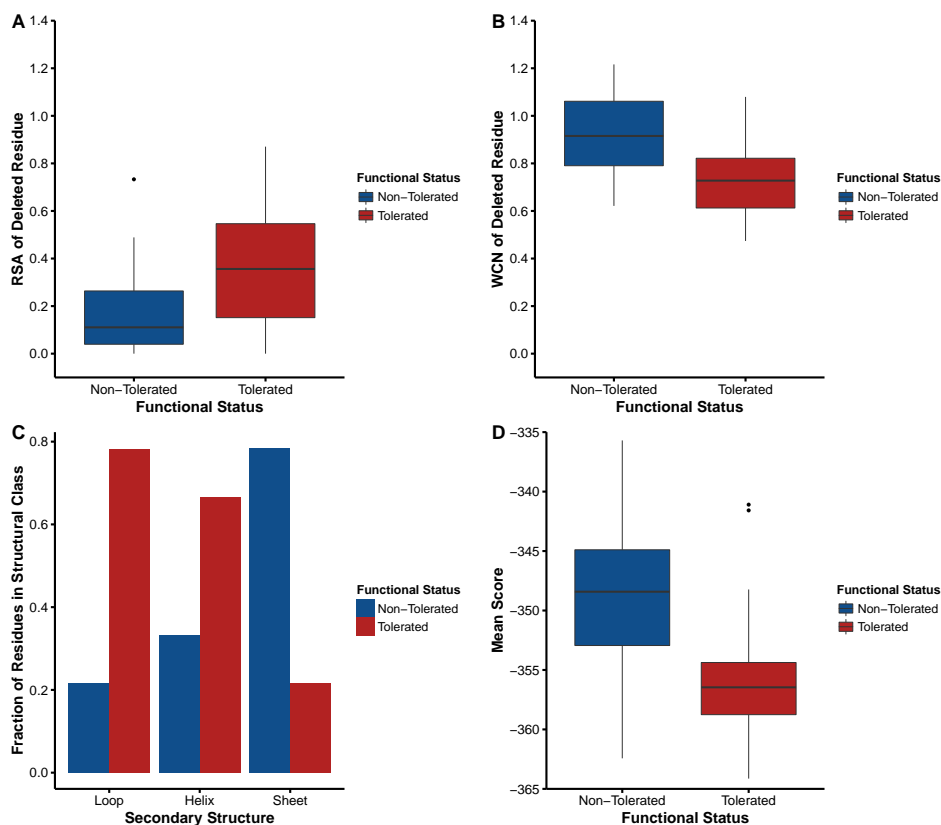


Figure 4.3: Distributions of Structural Properties for EGFP Mutants. (A) Distribution of RSA. Residues with tolerated deletions are more exposed than residues with non-tolerated deletions ( $t$  test:  $P = 1.030 \times 10^{-3}$ ). (B) Distribution of WCN. On average, residues with tolerated deletions have lower WCN than residues with non-tolerated deletions. ( $t$  test:  $P = 2.998 \times 10^{-7}$ ). (C) Distribution of mean score for EGFP Mutants. Residues that are tolerant to deletion have lower scores (i.e., more negative) on average than non-tolerant residues ( $t$  test:  $P = 2.084 \times 10^{-6}$ ). (D) Secondary structure of mutants. Non-tolerated deletions are colored in blue and tolerated deletions are in red. The majority of the residues deleted in the loop regions and alpha helix regions are tolerated and result in a functioning fluorescent phenotype. 78.3% and 66.7% of deleted residues are tolerated in loop and helical regions, respectively. However, only a small fraction of residues (21.6%) deleted in areas of the proteins that make up a beta sheet are tolerated.

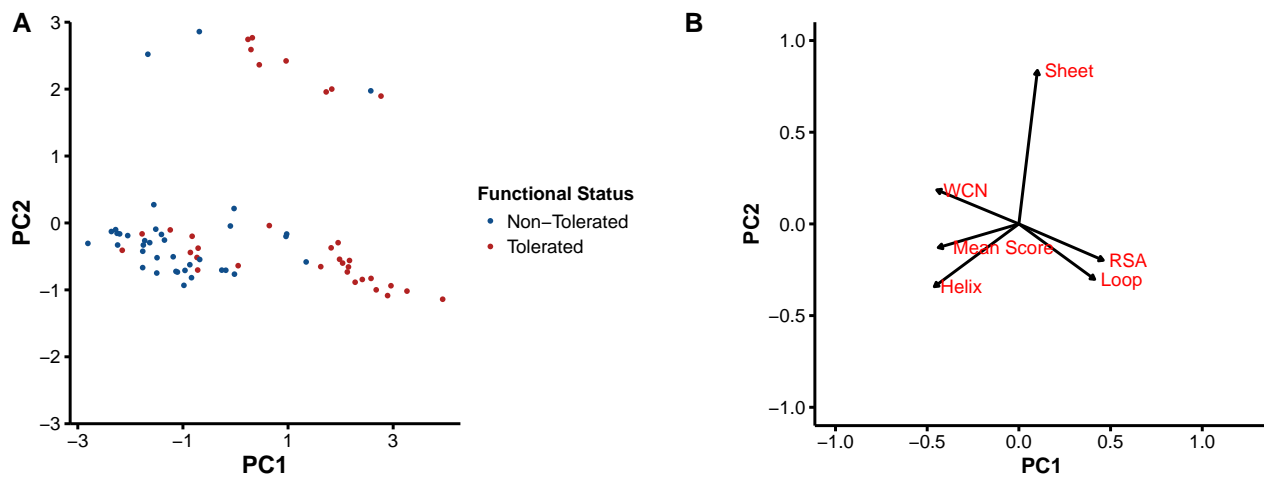


Figure 4.4: Comparison of Data Along Principal Component 1 versus Principal Component 2. A) Plot of PC1 vs. PC2. Data points are colored according to functional status. Functional mutants are blue and non-functional mutants are in red. Mutants are largely separated along PC1. B) Loadings of structural properties along principal component axes PC1 and PC2. Most structural properties contribute to PC1 except for beta sheet which is mostly loaded onto PC2.

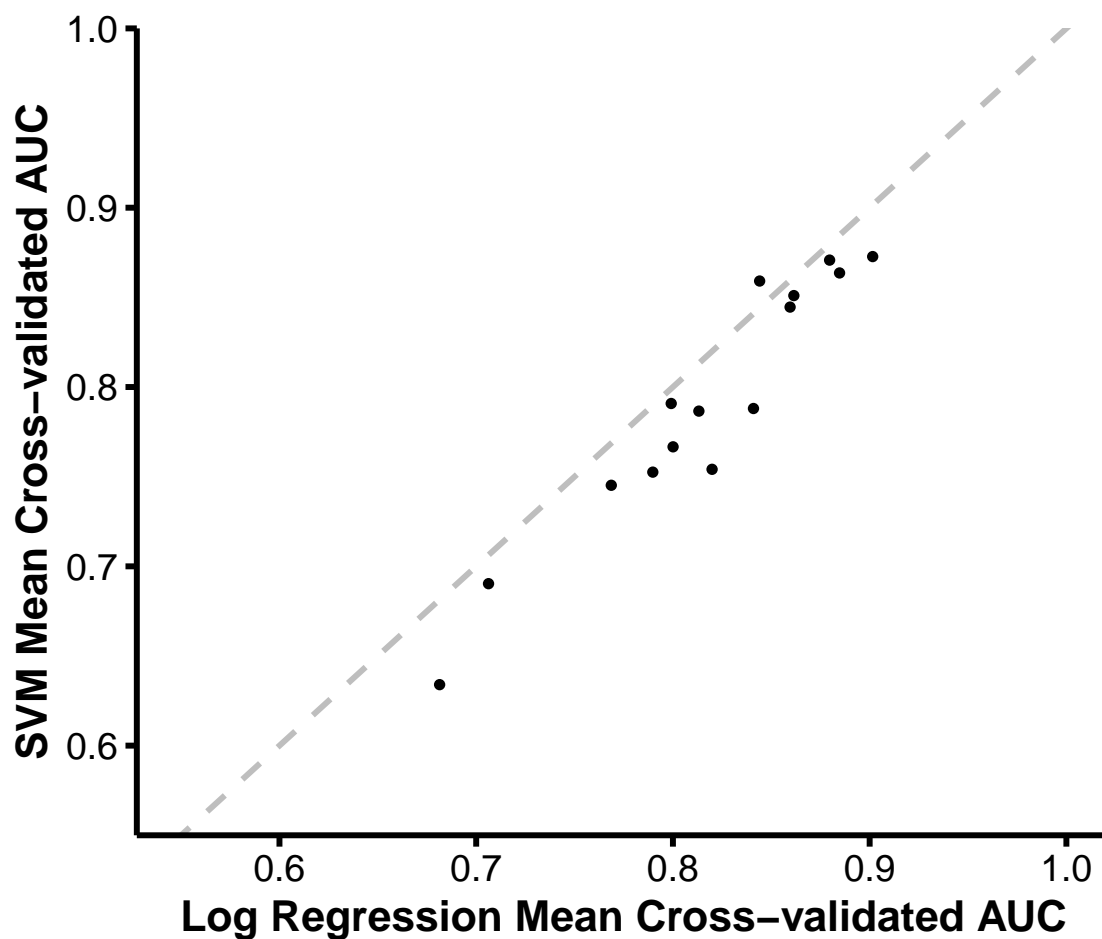


Figure 4.5: Comparisons of mean cross-validated AUC from SVM and Logistic Regression Models. For each model that has the exact same predictors the cross-validated AUC value from the SVM is plotted against the cross-validated AUC value from the logistic model. The dotted gray line represents the line  $y = x$ . For all but one model, logistic regression models with the same predictors have higher mean cross-validated AUC values.

## 4.6 Tables

Table 4.1: Summary of AUC values for logistic regression models using structure to predict functional status. The structural properties analyzed are RSA, WCN, SS, and mean score. The SS of a residue was classified as beta sheet, alpha helix or loop. Mean score is the mean of the Rosetta scores for 100 models of a given mutant. Each property was used as a single predictor or in combination with the other three structural predictors to predict the functional status of a given mutant. Functional status was taken from Arpino *et al.* [5]. We report the mean Area Under the Curve (AUC) of 100 trials for each model for the test data (mean cross-validated AUC). We also report the AUC for the model fitted on the entire data set (AUC of Model). Models are sorted in decreasing order by mean cross-validated AUC. The model with RSA, WCN and mean score has the best predictive ability.

Model	AUC of Model	Mean cross-validated AUC $\pm$ Standard Error
RSA + WCN + Mean Score	0.930	$0.902 \pm 0.0007$
RSA + WCN + SS	0.920	$0.885 \pm 0.0009$
RSA + WCN + Mean Score + SS	0.923	$0.880 \pm 0.0010$
WCN + Mean Score	0.875	$0.861 \pm 0.0004$
RSA + WCN	0.876	$0.860 \pm 0.0006$
WCN + SS + Mean Score	0.842	$0.844 \pm 0.0008$
WCN + SS	0.872	$0.841 \pm 0.0008$
WCN	0.817	$0.820 \pm 0.0005$
RSA + Mean Score	0.834	$0.814 \pm 0.0007$
Mean Score	0.875	$0.800 \pm 0.0005$
RSA + SS + Mean Score	0.850	$0.799 \pm 0.0012$
SS + Mean Score	0.842	$0.790 \pm 0.0014$
RSA + SS	0.808	$0.769 \pm 0.0011$
SS	0.844	$0.706 \pm 0.0020$
RSA	0.699	$0.681 \pm 0.0007$

Table 4.2: Summary of AUC values when using a support vector machine to predict functional status. The structural properties analyzed are RSA, WCN, SS, and mean score. Mean score is the mean of the Rosetta scores for 100 models of a given mutant. Each property was used as a single predictor or in combination with the other three structural predictors to predict the functional status of a given mutant. We report the mean Area Under the Curve (AUC) of 100 trials for each model for the test data (cross-validated AUC). We also report the AUC of the model for the model fitted on the entire data set (Model AUC). Models are sorted in decreasing order by mean cross-validated AUC. The model that is the best at making predictions is the model with RSA, WCN and mean score as structural predictors.

Model	Model AUC	Mean cross-validated AUC $\pm$ Standard Error
RSA + WCN + Mean Score	0.937	$0.873 \pm 0.0011$
RSA + WCN + Mean Score + SS	0.937	$0.871 \pm 0.0010$
RSA + WCN + SS	0.918	$0.864 \pm 0.0015$
WCN + SS + Mean Score	0.901	$0.859 \pm 0.0008$
WCN + Mean Score	0.908	$0.851 \pm 0.0009$
RSA + WCN	0.913	$0.845 \pm 0.0014$
RSA + SS + Mean Score	0.885	$0.791 \pm 0.0016$
WCN + SS	0.868	$0.788 \pm 0.0016$
RSA + Mean Score	0.872	$0.787 \pm 0.0020$
Mean Score	0.815	$0.767 \pm 0.0014$
WCN	0.820	$0.754 \pm 0.0016$
SS + Mean Score	0.852	$0.753 \pm 0.0022$
RSA + SS	0.821	$0.745 \pm 0.0025$
SS	0.864	$0.690 \pm 0.0020$
RSA	0.755	$0.634 \pm 0.0025$

## Chapter 5

# Computational prediction of zoonotic transmission of Machupo Virus

### 5.1 Introduction

This work was previously published as part of a paper in the journal *Journal of Virology*.<sup>1</sup>

Host-switching occurs when a virus has the ability to “jump” from its reservoir host species into another species, such as humans. The ability of viruses to host-switch into humans has resulted in outbreaks that have led to the loss of numerous human lives. One such virus is Machupo, a New World Arenavirus. The virus family Arenaviridae contains at least 23 viruses and is split into two main, geographically distinct groups, the Old World Arenaviruses and the New World Arenaviruses [13]. The New World Arenaviruses reside in South America, due to their association with their South American rodent hosts. Several New World Arenaviruses have the ability to infect humans and cause hemorrhagic fever. Junín (JUNV), Sabia (SABV), Guanarito (GTOV),

---

<sup>1</sup>S. A. Kerr, E. L. Jackson, O. I. Lungu, A. G. Meyer, A. Demogines, A. D. Ellington, G. Georgiou, C. O. Wilke, and S. L. Sawyer. Computational and Functional Analysis of the Virus-Receptor Interface Reveals Host Range Trade-Offs in New World Arenavirus. *Journal of Virology*, 89:11643–11653, 2013. A.G. Meyer, O. I. Lungu helped design experiments. S. Sawyer and C. O. Wilke helped to design the project and write the manuscript.

and Machupo (MACV) virus cause Venezuelan, Argentinian, Brazilian, and Bolivian hemorrhagic fever, respectively [70]. Mortality rates for infected humans are between 15 – 30% [9, 13].

The native host of Machupo virus is *Calomys callosus* (*C. callosus*), the large vesper mouse, a South American rodent. However, Machupo virus through cross-species transmission, has the ability to infect humans. Infection by Machupo causes hemorrhagic fever in humans that can result in severe sickness and death [13]. Machupo virus has been responsible for several deaths as a result of sporadic outbreaks within South America. Understanding the molecular mechanisms of how Machupo Virus and other viruses are able to host-switch will aid us in predicting disease outbreaks and developing critical strategies for preventing morbidity and mortality.

The primary viral protein responsible for mediating the interaction between the host cell and the virus is glycoprotein 1 (GP1), through its binding of the host receptor protein [1, 80, 81]. In New World Arenaviruses the host receptor utilized for entry is the transferrin receptor 1 (TfR1), a ubiquitous protein that is involved in cell iron-uptake [4]. After binding by GP1, the viral glycoprotein 2 (GP2) initiates fusion of the viral and host membranes. Efficient use of TfR1 has been shown to be the most important determinant of whether Machupo virus will successfully infect a new host [81]. It has been shown that Machupo virus cannot use *Rattus norvegicus* (Brown Rat) and *Mus musculus* (mouse) TfR1 as an efficient receptors for entry [81]. However, human TfR1 and the *C.callosus* are both viable receptors. The ability of Machupo virus to



use a host receptor is indicative of which species it can infect. Therefore, a computational method that predicts receptor binding will allow us to predict cellular entry by Machupo.

Using homology modeling and protein docking, we developed a method to predict binding efficiency within the interface of the MACV GP1-TfR1 interaction. Using Modeller, a homology modeling protocol, we modeled various MACV GP1-TfR1 interactions. After modeling, we used RosettaDock [14] to dock the receptors to GP1 and then computationally assessed the binding efficiency of each receptor. We then compared our computational results with experimental entry data for these interactions. Using this method, we were able to computationally confirm the ability of MACV to efficiently utilize a given host receptor and, by proxy, infect a given host species. We found that our modeling pipeline could accurately recapitulate MACV host entry patterns. In addition, our pipeline was able to successfully discriminate between the entry pattern between human hTfR1 and human hTfR1 L212V, a SNP for which there is preliminary evidence that has shown that it provides protection from MACV *in vitro* [20]. All together our computational pipeline was able to generally predict host entry patterns in Machupo virus.

## 5.2 Materials and Methods

We developed a computational pipeline in order to computational predict the binding affinity in the MACV GP1-TfR1 system. First using the co-crystal structure of MACV GP1-hTfR1 as a template, homology modeling

was used to determine structures for several MACV GP1-TfR1 complexes. We used the software Modeller [28] to model eight MACV GP1-TfR1 complexes. Due to its important function in cellular iron-uptake, transferrin receptor 1 is highly conserved. For example, the sequence identity between the rat and human transferrin receptor 1 is approximately 76 percent. This high sequence identity allowed for efficient homology modeling.

The TfR1s for this study included five naturally occurring TfR1s: *Calomys callous* (*C. callous*), *Rattus norvegicus* (rat), *Mus musculus* (mouse), *Homo sapiens* (human) and human L212V. The human L212V receptor is identical to hTfR1 except that there is a valine at position 212. In addition, we modeled three additional chimeras that had amino-acid swaps in the critical binding region between MACV GP1 and TfR1. We called these three chimeras rat-short, rat-long, and mouse-human. The rat-short chimera is the rat TfR1 with a five residue swap (SNDIP to NGVYL) from *C. callosus*, corresponding to residues 207 to 212 in hTfR1. The rat-long chimera is the rat TfR1 with a ten residue swap (SGSNIDPVEA to ASNGVYLES), which includes the five residue from rat-short along with five additional amino acids. These residues correspond to residues 205 to 215 in hTfR1. The mouse-human chimera is the mouse TfR1 with a five amino acid swap (NLDP to RLVYL) from the human TfR1. These amino acids correspond to residues 208 to 212 in hTfR1. Figure 5.1 depicts an alignment of the TfR1s used in this study.

Modeller needs a sequence alignment between the template structure and the target sequence. The sequences for the human, rat, and mouse trans-

ferrin receptors were supplied by the Sawyer Lab at UT Austin (see Kerr *et al.* (2015) for details). For the mouse and rat receptors, we only used the sequence regions that aligned with the hTfR1 sequence from the co-crystal structure. We made the appropriate swaps in the rat, mouse and human sequences to create sequences for the rat-short, rat-long, mouse-human and human L212V receptors. MAFFT [47,48] was used to align the template and target sequences. The “auto” flag was given to the program to select the optimal alignment protocol.

The template for all models was the co-crystal structure for MACV GP1-hTfR1 (PDB ID: 3KAS) (Figure 5.2). All of the non-amino acid residues were removed and the structure was renumbered so that the numbering started from 1. This was done so that the structure would be compatible with the Rosetta protein-modeling suite that was used for the docking protocol. For each MACV GP1-TfR1, we made 100 models using the basic Modeller homology modeling protocol. We then used the loop-modeling protocol to re-model the loops for each modeled structure. Therefore our resulting dataset consisted of 100 MACV GP1-TfR1 complexes for each MACV GP1-TfR1 interaction.

Afterwards, each of the 100 complexes from Modeller was re-docked in RosettaDock [14]. We re-docked the complexes to refine the docked orientation of the new TfR1 relative to the MACV GP1. Because the relative docking orientation of the new MACV GP1-TfR1 may be different than MACV GP1-hTfR1, we used the rigid-body moves in RosettaDock to allow for backbone movements that may occur that help the new complex properly dock. For each

structure, we generated 100 docked complexes using RosettaDock. Therefore the final modeling set for each transferrin receptor consisted of 10,000 docked complexes. When comparing across species we compared the mean interface score of the top 10 models for each complex as a comparison metric for binding. Figure 5.3 shows an overview of the computational modeling protocol.

Before docking the template for docking must be prepacked in Rosetta. For prepacking, we used the docking prepack protocol in Rosetta with the following flags:

```
-database /path/to/rosetta/database  
-l pdb_list.txt #List of structures to prepack  
-docking:partners A_B  
-ex1  
-ex2aro  
-use_input_sc  
-out:file:fullatom  
-out:path:pdb ./output_pdbs/
```

For docking, we used the RosettaDock docking protocol with the following flags:

```
-database /path/to/rosetta/database  
-l pdb_list.txt #List of prepacked structures  
-partners A_B
```

```
-dock_pert 3 8  
-spin  
-ex1  
-ex2aro  
-use_input_sc  
-nstruct 100  
-out:file:scorefile score_filename.fasc  
-out:path:pdb ./output_pdbs/
```

We also compared the results of our pipeline to functional entry data from Kerr *et al.* (2015) [49] to assess the accuracy of our pipeline.

### 5.3 Results

The results from the modeling approach can be seen in Figure 5.4. The mean interface score of the top ten models for a given complex is a computational measure of binding affinity, with more negative scores indicating higher binding affinity. As mentioned earlier, mouse and rat are inefficient cellular receptors for MACV. This is captured in the computational results. Both rat and mouse TfR1 have higher (less negative) interface scores implying Machupo does not bind to these receptors efficiently. Human, *C. callosus*, mouse-human and rat-long all have similar binding interface scores. These binding scores are much more negative indicating better binding between the MACV GP1 and TfR1. Recall that the mouse-human mutant was created by swapping five amino acids from human TfR1 that reside in the portion that contacts MACV

GP1 into the mouse TfR1. The rat-short and rat-long chimeras were created by swapping five and ten amino acids from the *C. callosus* TfR1 interface region into the rat receptor, respectively. These minor changes were sufficient to gain better binding. This indicates that minor changes within the binding region of the receptor can result in a change from a receptor that cannot be bound by MACV to one that can.

Previous research has indicated the presence of a SNP (L212V hTfR1) naturally occurring in the human population that provides some protection from MACV *in vitro* [20]. We tested our ability to recapitulate this result by modeling hTfR1 L212V and comparing to the wildtype hTfR1. Indeed, according to our computational analysis, the L212V mutation does decrease the binding affinity between hTfR1 and MACV GP1 (Figure 5.4). The predicted binding affinity of this mutant is similar to that of both rat and mouse, two inefficient receptors. This one SNP is sufficient to change the mean binding interface score from -8.855 to -6.980. This result recapitulates earlier work that found that this SNP provides some protection from MACV entry *in vitro*.

Interestingly, this same TfR1 SNP (L212V) has been shown to result in increased entry by Sabia and Junín, two related viruses [49]. Therefore this suggests that the L212V mutation has a unique structural effect in Machupo virus as compared to other related New World Arenaviruses. Examination of our modeled structures illuminated the structural ramifications of this SNP. MACV GP1, as compared to the other New World Arenavirus GP1s, has an extra looped region that contacts TfR1 (Figure 5.5). When mutating leucine

to valine at residue 212 in hTfR1, the interaction between hTfR1 and MACV GP1 in this loop is modified. This modified interaction appears to result in decreased binding between hTfR1 and MACV GP1. Therefore this loop serves a critical role in the binding of hTfR1 and MACV GP1. The N-terminus region of other pathogenic New World Arenaviruses (JUNV, CHPV, and SABV) is different. This region is shorter in these viruses and therefore these viruses do not have this looped motif.

## 5.4 Discussion

Machupo virus (MACV) is a New World Arenavirus that infects *C. callosus*, a rodent found in Bolivia. However, MACV has developed the ability to host jump into humans. Humans infected with MACV develop Bolivian Hemorrhagic Fever which can cause serious illness and death. Key mutations within the region of the host transferrin receptor 1 (hTfR1) can determine whether MACV can utilize a given receptor for cellular entry. Here we used homology modeling and protein-protein docking to develop a computational pipeline to predict efficient receptor use in MACV. Using our method, we were able separate receptors that can be used for MACV entry from those that cannot by accurately recapitulating experimental entry assays for MACV GP1 [49]. Both rat and mouse were shown to be inefficient receptors for MACV GP1 entry [81] and the results of our computational pipeline support this. Both rat and mouse TfR1 had more positive protein binding scores than both human and *C. callosus*. This indicates that these receptors result in much

less inefficient entry.

In addition, swapping key residues from the *C. callosus* TfR1 receptor interaction interface into the rat receptor resulted in more negative interaction scores. This suggests that these residues are critical for binding and swapping these residues into the rat receptor can result in increased binding between this receptor and MACV GP1. This finding was been supported *in vitro* [49]. Our method was also able to support a previous assertion that a human SNP, L212V, provides some protection against MACV GP1 *in vitro*. Our model containing this SNP had a much lower mean interface score as compared to the human hTfR1, indicating decreased binding. The totality of these results indicates that this homology modeling and docking protocol has the ability to recapitulate experimental data within the MACV GP1-TfR1 system.

Our method relies on the ability to accurately model the sequence of a target homologous protein on the template structure. The accuracy of this procedure of is highly dependent on the level of sequence identity between the template and the target sequences. In general, traditional approaches only create accurate models for proteins with a high sequence identity [106] with the template structure sequence. To study protein-protein interactions one needs highly accurate models that can be used for docking. Highly accurate models require an identity of over 50 percent [60]. There is high identity (approximately 70%) between the hTfR1 sequence and all the TfR1 target sequences used in this study. However, the New World Arenavirus GP1s are highly diverged. The sequence identities between Machupo and Junín, Sabia, and



Chapare GP1 are 47, 27, and 30, respectively [10]. Therefore modeling these viruses using the MACV GP1 as a template was extremely difficult and, as such, we could not model mutations on the viral side with this current protocol. Key regions of particular trouble are unstructured long loop regions. When modeling these GP1s with MACV GP1 as a template, this region might need to be modeled using advanced loop-modeling methods to ensure an accurate model.

Molecular Dynamics (MD) has been used to refine *de novo* Rosetta models. Lindert *et al.* [58] found that cycling MD and Rosetta resulted in models that had a lower root-mean-square deviation (RMSD) to the native structure as compared to models that were not cycled. Cycling both homology modeling and MD within the same protocol might result in the ability to model structures with low sequence homology. Adding both advanced loop-modeling techniques and MD to the current protocol might result in improved protein-protein binding predictions, particularly on the viral side. Understanding the effects of mutations within the virus and host proteins will help us locate key residues that allow for viral entry and the ability to computationally predict the effects of viral protein binding will help us screen for mutants of interest. Further development of methods will allow us to better understand how MACV and other New World Arenaviruses enter and infect cells.

## 5.5 Figures

### Natural Sequences

R.norvegicus(Rat)	VTI-NSGSNI-DPVEAPEG
C.callosus	VTIINASNGV-YLLESPAG
M.Musculus(Mouse)	VTIVQSNGNL-DPVESPEG
H.sapiens(Human)	VIIIVDKNGRLVYLVENPGG
Human L212V	VIIIVDKNGRLVYVVENPGG

### Tested Variants

Rat-Short	VTI-NSGNGV-YLVEAPEG
Rat-Long	VTI-NASNGV-YLLESPAG
Mouse-Human	VTIVQSNGRLVYLVESPEG

205215

Figure 5.1: Alignment of TfR1 receptors from various species along with the tested chimeras. The sequence numbering corresponds to the amino-acid position in the hTfR1 sequence. The five naturally occurring receptors included in this study are: *R. norvegicus*, *M. musculus*, *C. callosus* (the native host of Machupo virus), *H. sapiens*, and the *H. sapiens* L212V variant. The rat-short chimera is the rat TfR1 with a five residue swap from *C. callosus* which is indicated in red. The rat-long chimera is the rat TfR1 with a ten residue swap which includes the five residues from rat-long along with five additional amino acids (colored in blue). The mouse-human chimera is the mouse TfR1 with a five amino acid swap from the human TfR1. These amino acids are colored in green. The human L212V is identical to the hTfR1 except that there is a valine at position 212. This valine is colored in magenta in the hTfR1 L212V sequence.

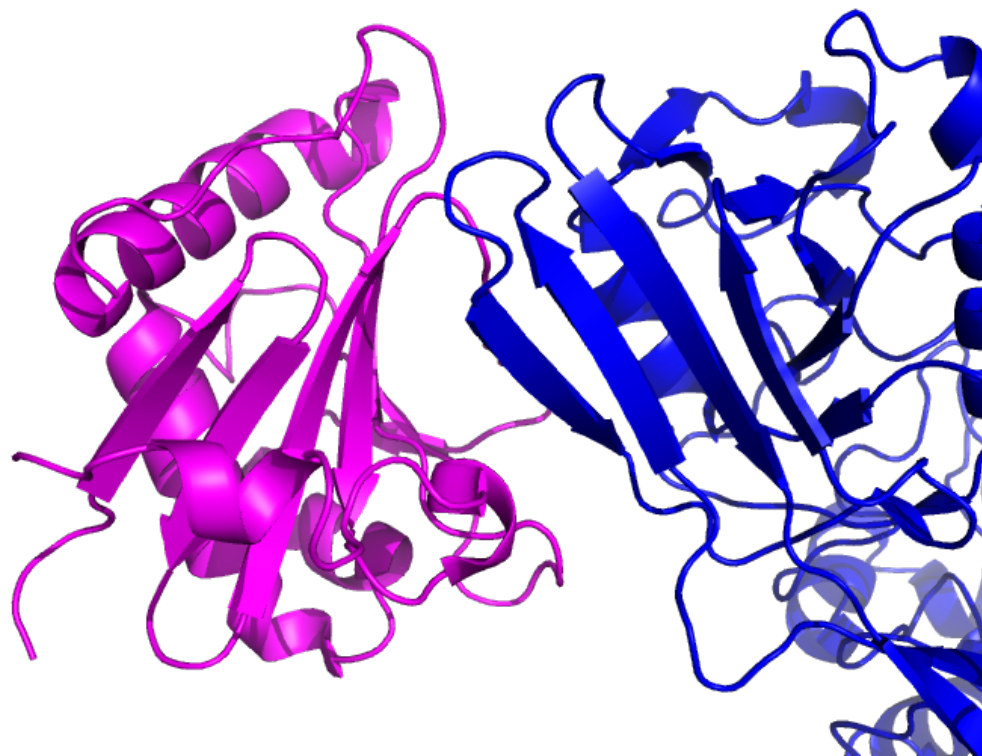


Figure 5.2: Interaction between human transferrin receptor 1 (hTfR1) and the Machupo Virus Glycoprotein 1 (MACV GP1). hTR1 is colored blue and MACV GP1 is colored blue. This is a visualization of the interaction between the apical domain of human TfR1 and MACV GP1. MACV uses its GP1 to bind to the TfR1 on the surface of the host's cell by interacting with the apical domain of TfR1.

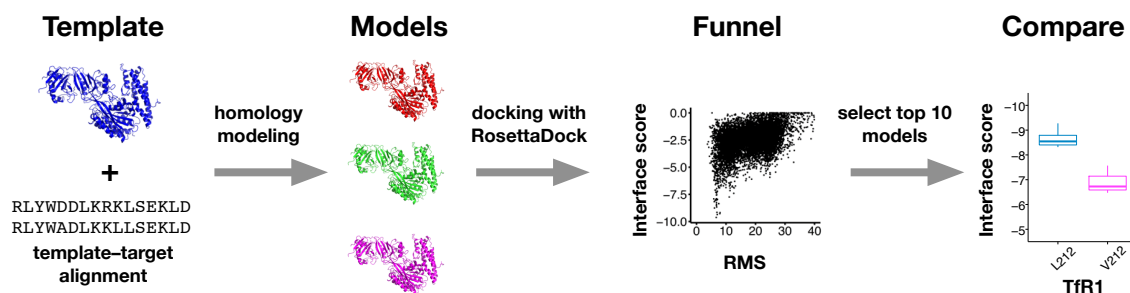


Figure 5.3: Computational Pipeline Overview. For each MACV GP1-TfR1, the target TfR1 sequence was aligned to the hTfR1 structure before modeling. After modeling each protein complex in Modeller, the complexes were re-docked using RosettaDock. Convergence of the docking protocol was assessed by plotting a RMS versus Interface Score plot and checking for a funnel. The mean interface score for the top ten scoring models for each MACV GP1-TfR1 complex was used as the proxy for binding affinity in subsequent analyses.

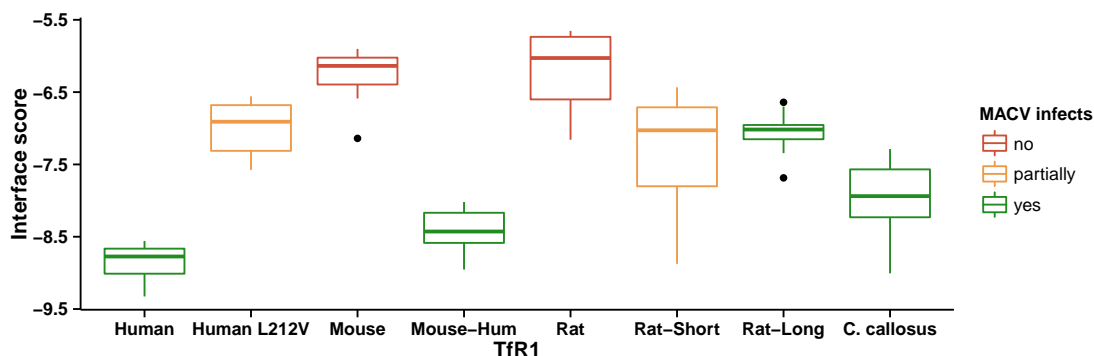


Figure 5.4: Mean Interface Scores for modeled MACV GP1-TfR1 Complexes. Each boxplot represents the distribution of the top ten scoring models for each TfR1 by interface score. Each boxplot is colored according to known infectivity information. Green coloring indicates efficient TfR1 receptors for entry. Yellow indicates receptors that are partially efficient. Red coloring indicates receptors that cannot be used as efficient receptors for entry. Overall, inefficient receptors have less negative interface scores indicating that binding is not as effective in those models. The human L212V model also has a much less negative average interface score as compared to the human model. This is consistent with experimental results suggesting that this SNP provides some protection from MACV *in vitro*.

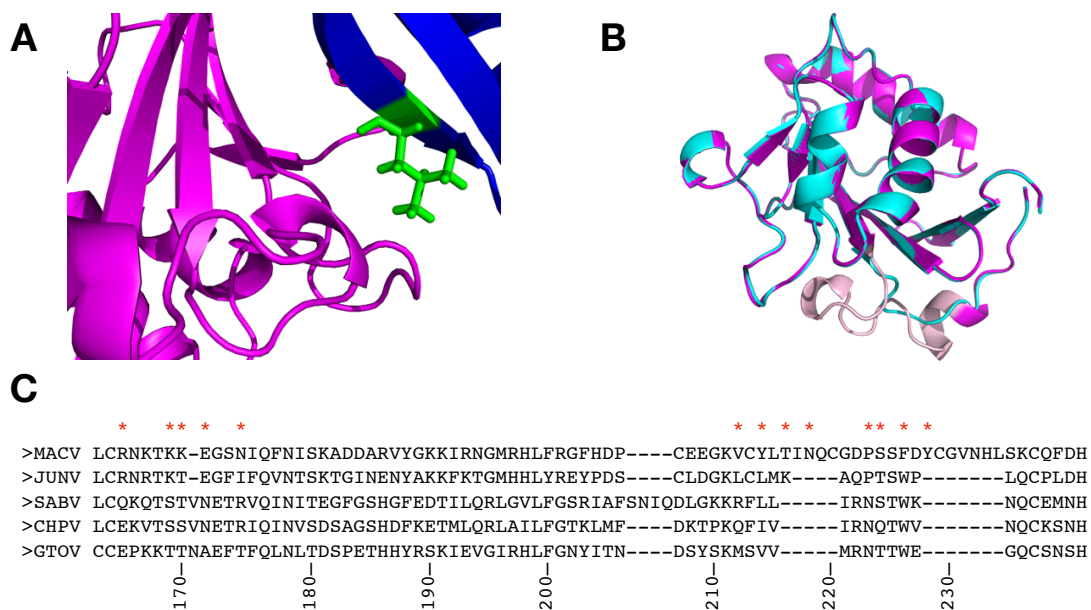


Figure 5.5: Interface between MACV GP1 and hTfR1. (A) Complex between MACV GP1-hTfR1 L212V. MACV GP1 is colored in magenta and the hTfR1 is colored blue. The L212V mutant is colored in green. (B) Superposition of a modeled JUNV GP1 onto the MACV GP1. JUNV GP1 is colored in cyan and MACV GP1 is colored in magenta. The loop region unique to MACV is colored pink. The JUNV GP1 is rotated relative to the position of the MACV GP1. (C) Alignment of the GP1 sequences of Machupo, Junín, Sabia, Chapare, and Guanarito. Red asterisks highlight residues that contact hTfR1 in MACV. Interaction residues between hTfR1 and MACV GP1 were calculated using FoldX [86].

## Chapter 6

### Conclusion

#### 6.1 Discussion

There have been several studies that have investigated the effect of biophysical properties on evolution at sites in proteins. There have also been several advances in our ability to model individual proteins and their interactions using protein design, homology modeling, and protein-protein docking. The development of such methods can be seen as the ultimate test of our understanding of how protein structure affects function. This dissertation work uses these computational modeling techniques to better understand how natural proteins evolve. The result is a better understanding of how these approaches can be used to understand natural proteins and has resulted in several key insights.

First, I performed a systematic comparison between natural and designed proteins. I found that in designed proteins hydrophobic residues were often underrepresented in the protein core. I also found that the relationship between solvent accessibility and site-wise variability was skewed in designed proteins. In natural proteins, it has been found that there is a significant, positive relationship between RSA and site variability with residues that are

more exposed to solvent exhibiting more variability. However, in designed proteins, residues on the surface were too conserved. Likewise, residues in the core were more variable than expected. In addition, I found that an intermediate amount of backbone flexibility during design resulted in sequences that were most similar to those observed in nature. Next I used protein design to predict evolution at individual sites in a protein. I found that protein design has limited utility in the prediction of how rapidly or slowly a given site will evolve within a protein. The amount of exposure to solvent, measured by RSA, and how densely packed a given residue is, measured by WCN, are much more successful at describing the variation we see at individual sites within proteins. These two studies highlight our need to develop better scoring functions and/or better sequence space search algorithms. Any further improvements in these two areas will improve our ability to design proteins.

In the third chapter, I examined the ability of protein modeling to predict the functional consequence of deletions. Although there have been numerous studies that have determined the ability of structure to predict variation at sites within proteins, most studies have focused predicting substitutions at sites. I systematically studied the effect of deletions on functional status in enhanced GFP (EGFP) and determined whether structure could be used to predict the functional repercussions of deletions in EGFP. I found that, in EGFP, the functional status of mutants with deletions follow patterns that are seen in studies on amino-acid substitutions. Residues that are more exposed to solvent are more tolerant to deletion. This is analogous to previous



studies that have shown that residues that are on the surface of proteins exhibit more site variability and evolve faster than residues within the core of a protein [12, 31, 32, 38, 65, 69, 85].

In addition, I found that residues that are densely packed, as measured by WCN, are less tolerant to deletion. This makes sense in light that it has been found that residues that are densely packed evolve slower [42, 55, 111, 112]. Although this work was only done on one protein, these results suggest that solvent exposure and local packing density, two well-studied quantities in terms of substitutions at sites, also have some effect on the functional consequences of deletions at sites in proteins. More experimental data on the functional status of individual deletions in other protein systems will be critical for future studies of structure and its effects on tolerance to deletions. In addition, I found that while using computational modeling approaches on their own to predict deletion tolerance was less effective than WCN, a much simpler structural quantity, adding the scores from computational modeling to a model with other structural predictors does increase predictive power. Lastly, this study further solidifies the role that protein contacts play in constraining the evolution of amino acids at sites. As in other studies [42, 55, 111], WCN seems to be the best predictor for predicting evolutionary change. Further study into mechanistic explanations relating WCN and other structural quantities would further our understanding of how structure constrains the evolution of natural proteins.

Lastly, I used computational modeling techniques to predict virus-host

interactions. I developed a pipeline that was able to recapitulate observed virus-host entry patterns in the MACV GP1–TfR1 system. The ability to computationally predict viral entry will provide an inexpensive, high throughput, and efficient way to study viral zoonosis and ultimately assess the risk of virus outbreaks to humans. However, although I could recapitulate virus-host patterns for MACV, I did discover some limitations to our approach. I was unable to create accurate models of the related New World Arenavirus GP1s due to the lack of sequence identity between the template used for modeling (MACV GP1) and the other viral GP1 proteins. Further study into the development of techniques that allow for the accurate modeling of targets that have low identity with the template structure will aid in refining this approach. In addition, further work on global docking of unbound protein partners will also help expand our ability to study other viral systems. Computational studies rely the use of structural data of protein-protein interactions. The ability to accurately determine native interactions will reduce our need to rely on difficult experimental techniques to obtain protein-protein complexes like X-ray crystallography.

## Bibliography

- [1] Jonathan Abraham, Kevin D. Corbett, Michael Farzan, Hyeryun Choe, and Stephen C. Harrison. Structural basis for receptor recognition by New World hemorrhagic fever arenaviruses. *Nature Structural & Molecular Biology*, 17(4):438–444, April 2010.
- [2] Andrew B. Allison, Dennis J. Kohler, Alicia Ortega, Elizabeth A. Hoover, Daniel M. Grove, Edward C. Holmes, and Colin R. Parrish. Host-Specific Parvovirus Evolution in Nature Is Recapitulated by In Vitro Adaptation to Different Carnivore Species. *PLoS Pathog*, 10(11):e1004475, November 2014.
- [3] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.
- [4] Nancy C. Andrews. Iron homeostasis: insights from genetics and animal models. *Nature Reviews Genetics*, 1(3):208–217, December 2000.
- [5] James A. J. Arpino, Samuel C. Reddington, Lisa M. Halliwell, Pierre J. Rizkallah, and D. Dafydd Jones. Random single amino acid deletion

sampling unveils structural tolerance and the benefits of helical registry shift on gfp folding and structure. *Structure*, 22:889–898, June 2014.

- [6] James A. J. Arpino, Pierre J. Rizkallah, and D. Dafydd Jones. Crystal Structure of Enhanced Green Fluorescent Protein to 1.35 Å Resolution Reveals Alternative Conformations for Glu222. *PLoS ONE*, 7(10):e47132, October 2012.
- [7] Ugo Bastolla, Markus Porto, H. Eduardo Roman, and Michele Vendruscolo. Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins: Structure, Function, and Bioinformatics*, 58(1):22–30, January 2005.
- [8] Jesse D. Bloom, Sy T. Labthavikul, Christopher R. Otey, and Frances H. Arnold. Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences*, 103(15):5869–5874, April 2006.
- [9] Luciana Borio, Thomas Inglesby, C. J. Peters, Alan L. Schmaljohn, James M. Hughes, Peter B. Jahrling, Thomas Ksiazek, Johnson Karl M., Meyerhoff. Andrea, Tara O’Toole, Michael S. Ascher, John Bartlett, Joel G. Breman, Edward M. Eitzen, Jr, Margaret Hamburg, Jerry Hauer, D. A. Henderson, Richard T. Johnson, Gigi Kwik, Marci Layton, Scott Lillibridge, Gary J. Nabel, Michael T. Osterholm, Trish M. Perl, Philip Russell, and Kevin Tonat. Hemorrhagic fever viruses as biological weapons: Medical and public health management. *JAMA*, 287(18):2391–2405, May 2002.

- [10] Thomas A. Bowden, Max Crispin, Stephen C. Graham, David J. Harvey, Jonathan M. Grimes, E. Yvonne Jones, and David I. Stuart. Unusual Molecular Architecture of the Machupo Virus Attachment Glycoprotein. *Journal of Virology*, 83(16):8259–8265, August 2009.
- [11] Daniel W. A. Buchan, Federico Minneci, Tim C. O. Nugent, Kevin Bryson, and David T. Jones. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research*, 41(W1):W349–W357, July 2013.
- [12] Carlos D. Bustamante, Jeffrey P. Townsend, and Daniel L. Hartl. Solvent Accessibility and Purifying Selection Within Proteins of *Escherichia coli* and *Salmonella enterica*. *Molecular Biology and Evolution*, 17(2):301–308, February 2000.
- [13] Rémi N. Charrel and Xavier de Lamballerie. Arenaviruses other than Lassa virus. *Antiviral Research*, 57(12):89–100, January 2003.
- [14] Sidhartha Chaudhury, Monica Berrondo, Brian D. Weitzner, Pravin Muthu, Hannah Bergman, and Jeffrey J. Gray. Benchmarking and Analysis of Protein Docking Performance in Rosetta v3.2. *PLoS ONE*, 6(8):e22477, August 2011.
- [15] Nicole Chaux, Philipp W. Messer, and Peter F. Arndt. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evolutionary Biology*, 7:191, 2007.

- [16] Rong Chen, Li Li, and Zhiping Weng. ZDOCK: An initial-stage protein-docking algorithm. *Proteins: Structure, Function, and Bioinformatics*, 52(1):80–87, July 2003.
- [17] Cyrus Chothia and Alexei V. Finkelstein. The Classification and Origins of Protein Folding Patterns. *Annual Review of Biochemistry*, 59(1):1007–1035, 1990.
- [18] Gavin C. Conant and Peter F. Stadler. Solvent Exposure Imparts Similar Selective Pressures across a Range of Yeast Proteins. *Molecular Biology and Evolution*, 26(5):1155–1161, May 2009.
- [19] Patrick Conway, Michael D. Tyka, Frank DiMaio, David E. Konerding, and David Baker. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Science*, 23(1):47–55, January 2014.
- [20] Ann Demogines, Jonathan Abraham, Hyeryun Choe, Michael Farzan, and Sara L. Sawyer. Dual Host-Virus Arms Races Shape an Essential Housekeeping Protein. *PLoS Biol*, 11(5):e1001571, May 2013.
- [21] Johan Desmet, Marc De Maeyer, Bart Hazes, and Ignace Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356(6369):539–542, April 1992.
- [22] D. Allan Drummond, Jesse D. Bloom, Christoph Adami, Claus O. Wilke, and Frances H. Arnold. Why highly expressed proteins evolve slowly.

- Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14338–14343, October 2005.
- [23] D. Allan Drummond and Claus O. Wilke. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell*, 134(2):341–352, July 2008.
  - [24] Julian Echave, Eleisha L. Jackson, and Claus O. Wilke. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *Physical Biology*, 12(2):025002, April 2015.
  - [25] Julian Echave, Stephanie J. Spielman, and Claus O. Wilke. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, 17(2):109–121, February 2016.
  - [26] Adrian H. Elcock. Prediction of functionally important residues based solely on the computed energetics of protein structure. *Journal of Molecular Biology*, 312(4):885–896, September 2001.
  - [27] Adrian H. Elcock and J. Andrew McCammon. Identification of protein oligomerization states by analysis of interface conservation. *Proceedings of the National Academy of Sciences*, 98(6):2990–2994, March 2001.
  - [28] András Fiser and Andrej Šali. Modeller: Generation and Refinement of Homology-Based Protein Structure Models. In *Methods in Enzymology*, volume 374 of *Macromolecular Crystallography, Part D*, pages 461–491. Academic Press, 2003.

- [29] Sarel J. Fleishman, Timothy A. Whitehead, Damian C. Ekiert, Cyrille Dreyfus, Jacob E. Corn, Eva-Maria Strauch, Ian A. Wilson, and David Baker. Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. *Science*, 332(6031):816–821, May 2011.
- [30] Mathieu Fourment and Mark J Gibbs. PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. *BMC Evolutionary Biology*, 6(1):1, January 2006.
- [31] Eric A. Franzosa and Yu Xia. Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. *Molecular Biology and Evolution*, 26(10):2387–2395, October 2009.
- [32] Eric A. Franzosa and Yu Xia. Independent Effects of Protein Core Size and Expression on Residue-Level Structure-Evolution Relationships. *PLoS ONE*, 7(10):e46602, October 2012.
- [33] Hunter B. Fraser, Aaron E. Hirsh, Lars M. Steinmetz, Curt Scharfe, and Marcus W. Feldman. Evolutionary Rate in the Protein Interaction Network. *Science*, 296(5568):750–752, April 2002.
- [34] Gregory D. Friedland, Nils-Alexander Lakomek, Christian Griesinger, Jens Meiler, and Tanja Kortemme. A Correspondence Between Solution-State Dynamics of an Individual Protein and the Sequence and Conformational Diversity of its Family. *PLoS Comput Biol*, 5(5):e1000393, May 2009.



- [35] Pablo Gainza, Kyle E. Roberts, and Bruce R. Donald. Protein Design Using Continuous Rotamers. *PLoS Comput Biol*, 8(1):e1002335, January 2012.
- [36] Pablo Gainza, Kyle E. Roberts, Ivelin Georgiev, Ryan H. Lilien, Daniel A. Keedy, Cheng-Yu Chen, Faisal Reza, Amy C. Anderson, David C. Richardson, Jane S. Richardson, and Bruce R. Donald. Chapter Five - osprey: Protein Design with Ensembles, Flexibility, and Provable Algorithms. In Amy E. Keating, editor, *Methods in Enzymology*, volume 523 of *Methods in Protein Design*, pages 87–107. Academic Press, 2013.
- [37] Ivelin Georgiev, Ryan H. Lilien, and Bruce R. Donald. Improved Pruning algorithms and Divide-and-Conquer strategies for Dead-End Elimination, with application to protein design. *Bioinformatics*, 22(14):e174–e183, July 2006.
- [38] Nick Goldman, Jeffrey L. Thorne, and David T. Jones. Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution. *Genetics*, 149(1):445–458, May 1998.
- [39] Jeffrey J. Gray, Stewart Moughon, Chu Wang, Ora Schueler-Furman, Brian Kuhlman, Carol A. Rohl, and David Baker. Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. *Journal of Molecular Biology*, 331(1):281–299, August 2003.

- [40] Mark A. Hallen, Daniel A. Keedy, and Bruce R. Donald. Dead-end elimination with perturbations (DEEPer): A provable protein design algorithm with continuous sidechain and backbone flexibility. *Proteins: Structure, Function, and Bioinformatics*, 81(1):18–39, January 2013.
- [41] Po-Ssu Huang, Yih-En Andrew Ban, Florian Richter, Ingemar Andre, Robert Vernon, William R. Schief, and David Baker. RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLoS ONE*, 6(8):e24109, August 2011.
- [42] Tsun-Tsao Huang, Mara L. del Valle Marcos, Jenn-Kang Hwang, and Julian Echave. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. *BMC Evolutionary Biology*, 14(1):78, April 2014.
- [43] Eleisha L. Jackson, Noah Ollikainen, Arthur W. Covert, Tanja Kortemme, and Claus O. Wilke. Amino-acid site variability among natural and designed proteins. *PeerJ*, 1:e211, November 2013.
- [44] D. Dafydd Jones. Triplet nucleotide removal at random positions in a target gene: the tolerance of TEM-1  $\beta$ -lactamase to an amino acid deletion. *Nucleic Acids Research*, 33(9):e80–e80, January 2005.
- [45] David T. Jones. Protein secondary structure prediction based on position-specific scoring matrices1. *Journal of Molecular Biology*, 292(2):195–202, September 1999.

- [46] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.
- [47] Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33(2):511–518, January 2005.
- [48] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, July 2002.
- [49] Scott A. Kerr, Eleisha L. Jackson, Oana I. Lungu, Austin G. Meyer, Ann Demogines, Andrew D. Ellington, George Georgiou, Claus O. Wilke, and Sara L. Sawyer. Computational and Functional Analysis of the Virus-Receptor Interface Reveals Host Range Trade-Offs in New World Arenaviruses. *Journal of Virology*, 89(22):11643–11653, November 2015.
- [50] Brian Kuhlman and David Baker. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences*, 97(19):10383–10388, September 2000.
- [51] Brian Kuhlman, Gautam Dantas, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science*, 302(5649):1364–1368, November 2003.

- [52] Themis Lazaridis and Martin Karplus. Effective energy function for proteins in solution. *Proteins: Structure, Function, and Bioinformatics*, 35(2):133–152, 1999.
- [53] Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis, Oliver F. Lange, James Thompson, Ron Jacak, Kristian W. Kaufman, P. Douglas Renfrew, Colin A. Smith, Will Sheffler, Ian W. Davis, Seth Cooper, Adrien Treuille, Daniel J. Mandell, Florian Richter, Yih-En Andrew Ban, Sarel J. Fleishman, Jacob E. Corn, David E. Kim, Sergey Lyskov, Monica Berrondo, Stuart Mentzer, Zoran Popovi, James J. Havranek, John Karanicolas, Rhiju Das, Jens Meiler, Tanja Kortemme, Jeffrey J. Gray, Brian Kuhlman, David Baker, and Philip Bradley. Chapter nineteen - Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. In Michael L. Johnson and Ludwig Brand, editor, *Methods in Enzymology*, volume 487 of *Computer Methods, Part C*, pages 545–574. Academic Press, 2011.
- [54] Emmanuel D. Levy, Subhajyoti De, and Sarah A. Teichmann. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proceedings of the National Academy of Sciences*, 109(50):20461–20466, December 2012.
- [55] H. Liao, W. Yeh, D. Chiang, R. L. Jernigan, and B. Lustig. Protein sequence entropy is closely related to packing density and hydrophobicity. *Protein Engineering Design and Selection*, 18(2):59–64, February 2005.

- [56] David A. Liberles, Sarah A. Teichmann, Ivet Bahar, Ugo Bastolla, Jesse Bloom, Erich Bornberg-Bauer, Lucy J. Colwell, A. P. Jason de Koning, Nikolay V. Dokholyan, Julian Echave, Arne Elofsson, Dietlind L. Gerloff, Richard A. Goldstein, Johan A. Grahnen, Mark T. Holder, Clemens Lakner, Nicholas Lartillot, Simon C. Lovell, Gavin Naylor, Tina Perica, David D. Pollock, Tal Pupko, Lynne Regan, Andrew Roger, Nimrod Rubinstein, Eugene Shakhnovich, Kimmen Sjlander, Shamil Sunyaev, Ashley I. Teufel, Jeffrey L. Thorne, Joseph W. Thornton, Daniel M. Weinreich, and Simon Whelan. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Science*, 21(6):769–785, June 2012.
- [57] Chih-Peng Lin, Shao-Wei Huang, Yan-Long Lai, Shih-Chung Yen, Chien-Hua Shih, Chih-Hao Lu, Cuen-Chao Huang, and Jenn-Kang Hwang. Deriving protein dynamical properties from weighted protein contact number. *Proteins: Structure, Function, and Bioinformatics*, 72(3):929–935, August 2008.
- [58] Steffen Lindert, Jens Meiler, and J. Andrew McCammon. Iterative molecular dynamics—rosetta protein structure refinement protocol to improve model quality. *Journal of Chemical Theory and Computation*, 9(8):3843–3847, August 2013.
- [59] Mara Laura Marcos and Julian Echave. Too packed to change: side-chain packing and site-specific substitution rates in protein evolution.

*PeerJ*, 3:e911, April 2015.

- [60] Marc A. Martí-Renom, Ashley C. Stuart, Andràs Fiser, Roberto Snchez, Francisco Melo, and Andrej ali. Comparative Protein Structure Modeling of Genes and Genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29(1):291–325, 2000.
- [61] Itay Mayrose, Dan Graur, Nir Ben-Tal, and Tal Pupko. Comparison of Site-Specific Rate-Inference Methods for Protein Sequences: Empirical Bayesian Methods Are Superior. *Molecular Biology and Evolution*, 21(9):1781–1791, September 2004.
- [62] Austin G. Meyer and Claus O. Wilke. Integrating Sequence Variation and Protein Structure to Identify Sites under Selection. *Molecular Biology and Evolution*, 30(1):36–44, January 2013.
- [63] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2015. R package version 1.6-7.
- [64] Julian Mintseris, Kevin Wiehe, Brian Pierce, Robert Anderson, Rong Chen, Jol Janin, and Zhiping Weng. Proteinprotein docking benchmark 2.0: An update. *Proteins: Structure, Function, and Bioinformatics*, 60(2):214–216, August 2005.

- [65] Leonid A. Mirny and Eugene I. Shakhnovich. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function1. *Journal of Molecular Biology*, 291(1):177–196, September 1999.
- [66] Susanne Moelbert, Eldon Emberly, and Chao Tang. Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. *Protein Science*, 13(3):752–762, March 2004.
- [67] Lucas Gregorio Nivón, Rocco Moretti, and David Baker. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLOS ONE*, 8(4):e59004, April 2013.
- [68] Noah Ollikainen and Tanja Kortemme. Computational Protein Design Quantifies Structural Constraints on Amino Acid Covariation. *PLoS Comput Biol*, 9(11):e1003313, November 2013.
- [69] John Overington, Dan Donnelly, Mark S. Johnson, Andrej Šali, and Tom L. Blundell. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Science*, 1(2):216–226, February 1992.
- [70] Slobodan Paessler and David H. Walker. Pathogenesis of the Viral Hemorrhagic Fevers. *Annual Review of Pathology: Mechanisms of Disease*, 8(1):411–440, 2013.

- [71] Kaifang Pang, Chao Cheng, Zhenyu Xuan, Huanye Sheng, and Xiaotu Ma. Understanding protein evolutionary rate by integrating gene co-expression with protein interactions. *BMC Systems Biology*, 4(1):179, December 2010.
- [72] Stefano Pascarella and Patrick Argos. Analysis of insertions/deletions in protein structures. *Journal of Molecular Biology*, 224(2):461–471, March 1992.
- [73] Osnat Penn, Eyal Privman, Giddy Landan, Dan Graur, and Tal Pupko. An Alignment Confidence Score Capturing Robustness to Guide Tree Uncertainty. *Molecular Biology and Evolution*, 27(8):1759–1767, August 2010.
- [74] Brian G. Pierce, Kevin Wiehe, Howook Hwang, Bong-Hyun Kim, Thom Vreven, and Zhiping Weng. ZDOCK server: interactive docking prediction of proteinprotein complexes and symmetric multimers. *Bioinformatics*, 30(12):1771–1773, June 2014.
- [75] Craig T. Porter, Gail J. Bartlett, and Janet M. Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, 32(suppl 1):D129–D133, January 2004.
- [76] Markus Porto, H. Eduardo Roman, Michele Vendruscolo, and Ugo Bastolla. Prediction of Site-Specific Amino Acid Distributions and Limits



- of Divergent Evolutionary Changes in Protein Sequences. *Molecular Biology and Evolution*, 22(3):630–638, March 2005.
- [77] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. Fast tree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE*, 5(3):e9490, March 2010.
- [78] Tal Pupko, Rachel E. Bell, Itay Mayrose, Fabian Glaser, and Nir Ben-Tal. Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18(suppl 1):S71–S77, July 2002.
- [79] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
- [80] Sheli R. Radoshitzky, Jonathan Abraham, Christina F. Spiropoulou, Jens H. Kuhn, Dan Nguyen, Wenhui Li, Jane Nagel, Paul J. Schmidt, Jack H. Nunberg, Nancy C. Andrews, Michael Farzan, and Hyeryun Choe. Transferrin receptor 1 is a cellular receptor for New World haemorrhagic fever arenaviruses. *Nature*, 446(7131):92–96, March 2007.
- [81] Sheli R. Radoshitzky, Jens H. Kuhn, Christina F. Spiropoulou, Csar G. Albario, Dan P. Nguyen, Jorge Salazar-Bravo, Tatyana Dorfman, Amy S. Lee, Enxiu Wang, Susan R. Ross, Hyeryun Choe, and Michael Farzan.

Receptor determinants of zoonotic transmission of New World hemorrhagic fever arenaviruses. *Proceedings of the National Academy of Sciences*, 105(7):2664–2669, February 2008.

- [82] Duncan C. Ramsey, Michael P. Scherrer, Tong Zhou, and Claus O. Wilke. The Relationship Between Relative Solvent Accessibility and Evolutionary Rate in Protein Evolution. *Genetics*, 188(2):479–488, June 2011.
- [83] Daniela Röthlisberger, Olga Khersonsky, Andrew M. Wollacott, Lin Jiang, Jason DeChancie, Jamie Betker, Jasmine L. Gallaher, Eric A. Althoff, Alexandre Zanghellini, Orly Dym, Shira Albeck, Kendall N. Houk, Dan S. Tawfik, and David Baker. Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192):190–195, May 2008.
- [84] Torsten Schaller, Karen E. Ocwieja, Jane Rasaiyaah, Amanda J. Price, Troy L. Brady, Shoshannah L. Roth, Stphane Hu, Adam J. Fletcher, KyeongEun Lee, Vineet N. KewalRamani, Mahdad Noursadeghi, Richard G. Jenner, Leo C. James, Frederic D. Bushman, and Greg J. Towers. HIV-1 Capsid-Cyclophilin Interactions Determine Nuclear Import Pathway, Integration Targeting and Replication Efficiency. *PLoS Pathog*, 7(12):e1002439, December 2011.
- [85] Michael P. Scherrer, Austin G. Meyer, and Claus O. Wilke. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evolutionary Biology*, 12(1):179, September 2012.

- [86] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. *Nucleic Acids Research*, 33(suppl 2):W382–W388, July 2005.
- [87] Adrian W. R. Serohijos, Zilvinas Rimas, and Eugene I. Shakhnovich. Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly. *Cell Reports*, 2(2):249–256, August 2012.
- [88] Amir Shahmoradi, Dariya K. Sydykova, Stephanie J. Spielman, Eleisha L. Jackson, Eric T. Dawson, Austin G. Meyer, and Claus O. Wilke. Predicting Evolutionary Site Variability from Structure in Viral Proteins: Buriedness, Packing, Flexibility, and Design. *Journal of Molecular Evolution*, 79(3-4):130–142, September 2014.
- [89] Chien-Hua Shih, Chih-Min Chang, Yeong-Shin Lin, Wei-Cheng Lo, and Jenn-Kang Hwang. Evolutionary information hidden in a single protein structure. *Proteins: Structure, Function, and Bioinformatics*, 80(6):1647–1657, June 2012.
- [90] Tobias Sikosek and Hue Sun Chan. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of The Royal Society Interface*, 11(100):20140419, November 2014.
- [91] Alan M. Simm, Amy J. Baldwin, Kathy Busse, and D. Dafydd Jones. Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1  $\beta$ -lactamase. *FEBS Letters*, 581(21):3904–3908, August 2007.

- [92] Manuela Sironi, Rachele Cagliani, Diego Forni, and Mario Clerici. Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nature Reviews Genetics*, 16(4):224–236, April 2015.
- [93] Colin A. Smith and Tanja Kortemme. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. *Journal of Molecular Biology*, 380(4):742–756, July 2008.
- [94] Colin A. Smith and Tanja Kortemme. Structure-Based Prediction of the Peptide Sequence Space Recognized by Natural and Synthetic PDZ Domains. *Journal of Molecular Biology*, 402(2):460–474, September 2010.
- [95] Stephanie J. Spielman, Eric T. Dawson, and Claus O. Wilke. Limited Utility of Residue Masking for Positive-Selection Inference. *Molecular Biology and Evolution*, 31(9):2496–2500, September 2014.
- [96] Stephanie J. Spielman and Claus O. Wilke. Pyvolve: A Flexible Python Module for Simulating Sequences along Phylogenies. *PLoS ONE*, 10(9):e0139047, September 2015.
- [97] Stephanie J. Spielman and Claus O. Wilke. The Relationship between dN/dS and Scaled Selection Coefficients. *Molecular Biology and Evolution*, 32(4):1097–1108, April 2015.

- [98] Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, November 2006.
- [99] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, May 2014.
- [100] Jeet Sukumaran and Mark T. Holder. DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571, June 2010.
- [101] Gian Gaetano Tartaglia, Sebastian Pechmann, Christopher M. Dobson, and Michele Vendruscolo. Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends in Biochemical Sciences*, 32(5):204–206, May 2007.
- [102] Martin S. Taylor, Chris P. Ponting, and Richard R. Copley. Occurrence and Consequences of Coding Sequence Insertions and Deletions in Mammalian Genomes. *Genome Research*, 14(4):555–566, April 2004.
- [103] Matthew Z. Tien, Austin G. Meyer, Dariya K. Sydykova, Stephanie J. Spielman, and Claus O. Wilke. Maximum Allowed Solvent Accessibilities of Residues in Proteins. *PLoS ONE*, 8(11):e80635, November 2013.
- [104] Nobuhiko Tokuriki, Christopher J. Oldfield, Vladimir N. Uversky, Igor N. Berezovsky, and Dan S. Tawfik. Do viral proteins possess unique bio-

- physical features? *Trends in Biochemical Sciences*, 34(2):53–59, February 2009.
- [105] gnes Tth-Petrczy and Dan S. Tawfik. Protein Insertions and Deletions Enabled by Neutral Roaming in Sequence Space. *Molecular Biology and Evolution*, 30(4):761–771, April 2013.
  - [106] Andrej Šali. Modelling mutations and homologous proteins. *Current Opinion in Biotechnology*, 6(4):437–451, 1995.
  - [107] William S. J. Valdar and Janet M. Thornton. Conservation helps to identify biologically relevant crystal contacts<sup>1</sup>. *Journal of Molecular Biology*, 313(2):399–416, October 2001.
  - [108] Larry Wasserman. *All of statistics: a concise course in statistical inference*. New York: Springer, 2004.
  - [109] Claus O. Wilke and D. Allan Drummond. Signatures of protein biophysics in coding sequence evolution. *Current Opinion in Structural Biology*, 20(3):385–389, June 2010.
  - [110] Jian-Rong Yang, Ben-Yang Liao, Shi-Mei Zhuang, and Jianzhi Zhang. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proceedings of the National Academy of Sciences*, 109(14):E831–E840, April 2012.
  - [111] So-Wei Yeh, Tsun-Tsao Huang, Jen-Wei Liu, Sung-Huan Yu, Chien-Hua Shih, Jenn-Kang Hwang, and Julian Echave. Local Packing Density

- Is the Main Structural Determinant of the Rate of Protein Sequence Evolution at Site Level. *BioMed Research International*, 2014:e572409, July 2014.
- [112] So-Wei Yeh, Jen-Wei Liu, Sung-Huan Yu, Chien-Hua Shih, Jenn-Kang Hwang, and Julian Echave. Site-Specific Structural Constraints on Protein Sequence Evolutionary Divergence: Local Packing Density versus Solvent Exposure. *Molecular Biology and Evolution*, 31(1):135–139, January 2014.
- [113] Golan Yona and Michael Levitt. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory1. *Journal of Molecular Biology*, 315(5):1257–1275, February 2002.
- [114] Ali Zarrinpar, Sang-Hyun Park, and Wendell A. Lim. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature*, 426(6967):676–680, December 2003.
- [115] Jianzhi Zhang and Jian-Rong Yang. Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, 16(7):409–420, July 2015.

## Vita

Eleisha Jackson graduated from Willow Canyon High School in Surprise, Arizona in 2008. Afterwards, she attended the University of Arizona in Tucson. She graduated from the University of Arizona in 2012 with a Bachelor of Science in Mathematics and minors in biology and art history. After graduation, she joined the Ecology, Evolution and Behavior program at the University of Texas at Austin in the fall of 2012.

Permanent address: [eleishaljackson@gmail.com](mailto:eleishaljackson@gmail.com)

This dissertation was typeset with L<sup>A</sup>T<sub>E</sub>X<sup>†</sup> by the author.

---

<sup>†</sup>L<sup>A</sup>T<sub>E</sub>X is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's T<sub>E</sub>X Program.